THE DEVELOPMENT OF ACCURATE FORCE FIELDS FOR PROTEIN SIMULATION

by

YUANFANG JIAO

B.S., Tianjin University, China, 2007

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Chemistry
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Abstract

Computer simulations have provided a wealth of information concerning a wide range of systems. The precision of computer simulation results depends on the degree of sampling (time scales) achieved, while the accuracy of the results (given sufficient sampling) depends on the quality of force field used. A force field provides a description of the energy for a system of interest. Recently, we have been developing a Kirkwood Buff (KB) force field for molecular dynamics simulations of biological systems. This force field is based on the KB Theory of solutions, emphasizing the accurate description of intermolecular interactions, and reasonably reproducing a range of other physical properties from experiment. In this approach simulation results in terms of KB integrals can be directly compared with experimental data through a KB analysis of the solution properties. The approach therefore provides a simple and clear method to test the capability of a force field. Here we firstly studied a series of alcohol-water mixtures in an attempt to validate the transferability and additivity of the force field. A general fluctuation theory was applied to investigate the properties of these systems, and to compare with computer simulation results. The possible effects of cosolvents on peptides and proteins were then investigated using N-methylacetamide as model for the peptide backbone and urea as cosolvent. A possible explanation for the urea denaturation of protein structure was provided using a thermodynamics point of view involving transfer free energies and preferential interactions obtained from the KB integrals. Finally, potentials for protein backbone and sidechain torsions were developed by fitting to quantum mechanical calculations and NMR data. Simulations of a variety of peptides and proteins in aqueous solutions were then performed to demonstrate the overall reliability of the force field.

THE DEVELOPMENT OF ACCURATE FORCE FIELDS FOR PROTEIN SIMULATION

by

YUANFANG JIAO

B.S., Tianjin University, China, 2007

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Chemistry
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Dr. Paul E. Smith

Abstract

Computer simulations have provided a wealth of information concerning a wide range of systems. The precision of computer simulation results depends on the degree of sampling (time scales) achieved, while the accuracy of the results (given sufficient sampling) depends on the quality of force field used. A force field provides a description of the energy for a system of interest. Recently, we have been developing a Kirkwood Buff (KB) force field for molecular dynamics simulations of biological systems. This force field is based on the KB Theory of solutions, emphasizing the accurate description of intermolecular interactions, and reasonably reproducing a range of other physical properties from experiment. In this approach simulation results in terms of KB integrals can be directly compared with experimental data through a KB analysis of the solution properties. The approach therefore provides a simple and clear method to test the capability of a force field. Here we firstly studied a series of alcohol-water mixtures in an attempt to validate the transferability and additivity of the force field. A general fluctuation theory was applied to investigate the properties of these systems, and to compare with computer simulation results. The possible effects of cosolvents on peptides and proteins were then investigated using N-methylacetamide as model for the peptide backbone and urea as cosolvent. A possible explanation for the urea denaturation of protein structure was provided using a thermodynamics point of view involving transfer free energies and preferential interactions obtained from the KB integrals. Finally, potentials for protein backbone and sidechain torsions were developed by fitting to quantum mechanical calculations and NMR data. Simulations of a variety of peptides and proteins in aqueous solutions were then performed to demonstrate the overall reliability of the force field.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my research advisor Dr. Paul E. Smith for his guidance, instruction, and support during my graduate study. I feel very lucky to have the opportunity to learn from such a brilliant chemist. The longer I work with him, the more I respect him. His rigorous study attitude and endless enthusiasm to science make a profound impression to me and motivate me to persist in future study.

I also appreciate Dr. Jianhan Chen for a lot of valuable advices and suggestions. His smart ideas always refresh my mind and provide me a new viewpoint of thinking about science.

To the Smith Group (Samantha Weerasinghe, Myungshim Kang, Feng Chen, Moon Bae Gee, Jin Zou, Elizabeth Ploetz, Shu Dai, Sadish Karunaweera, Mohomed Nawavi Mohomed Naleem, Gayani Pallewela), I have spent most of time with them in the lab. I want to thank them all to support me through these years and thanks for creating an atmosphere of curiousness to knowledge and caring for others in the lab.

To the members of Dr. Chen's group (Debabani Ganguly, Chester McDowell, Weihong Zhang, Jian Gao, Timothy Click), I really enjoy the Journal Club every week with them. I have a lot of fun to study and discuss with them.

Thanks to all my Ph.D. committee members Dr. Stefan Bossmann, Dr. Victor Chikan, Dr. Jianhan Chen and Dr. Jishu Shi for their valuable time and efforts.

Thanks to NIH for the funding to support our research.

Finally, I am grateful for my family for all their love, support and guidance through whole my life. It is them that have kept me going through all the difficulties.

Dedication

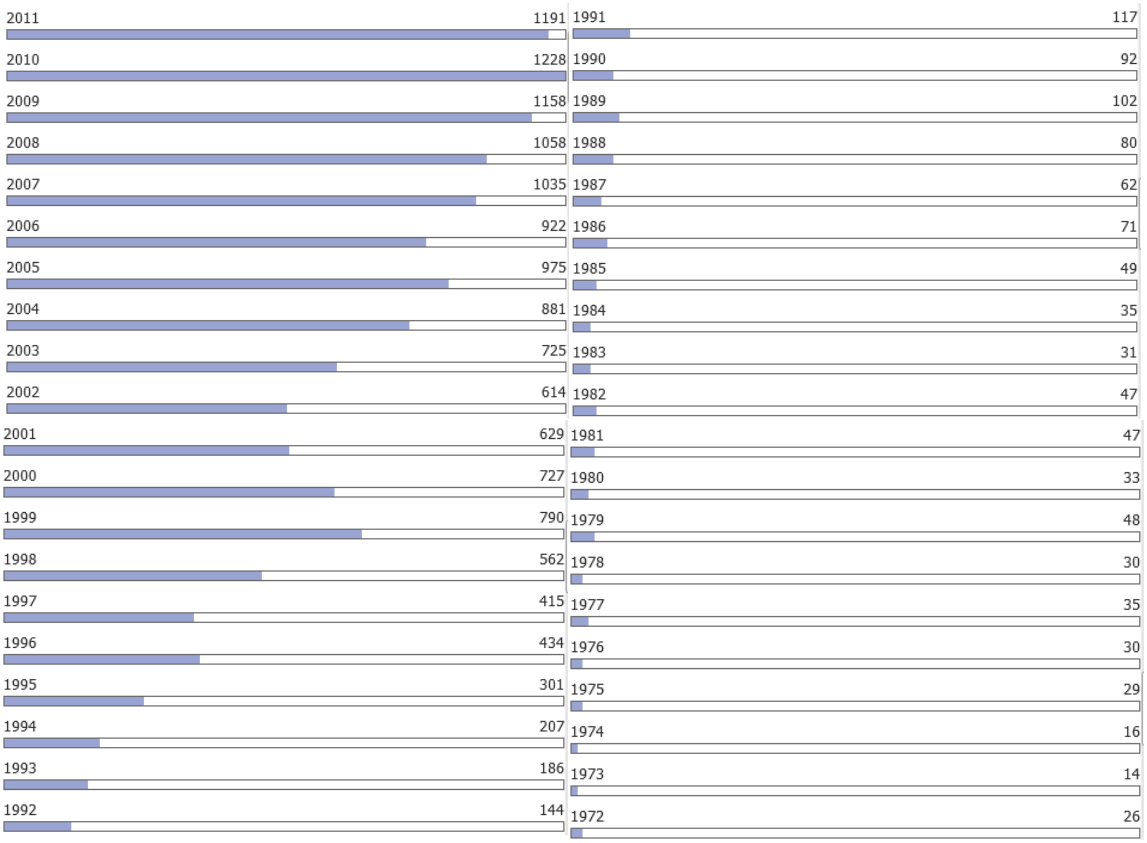
To my parents

Chapter 1 - Introduction

1.1 General Introduction

Since 1970, with the development of modern computer technology, computational chemistry has becomes an increasingly popular field. It is easy to see from Fig. 1.1 displaying the statistics analyzed by ACS database that the number of related publications increased at a dramatic speed during the latest two decades.

Figure 1.1 Numbers of publications about computational chemistry by year from 1972 to 2011 according to ACS database.



The goal of computational chemistry is not only to reproduce experimental results, but also to provide information that cannot be acquired from experiment. For instance, it is now accepted as a matter of fact that bio-macromolecules require some flexibility in order to perform many biological functions. This leads to a dynamical picture of protein behavior. The problem is

many experimental tools, like X-ray crystallography, can only provide us with relatively rigid equilibrium molecular structures. Under most circumstances experimental methods cannot reveal structural details about the conformational transitions of molecules in different environments. Fortunately, such information can, in principal, be directly accessed via computational approaches.

Nowadays, Computational Chemistry is widely applied in a variety of areas to facilitate a deeper understanding of experimental data. It can help crystallographers to determine and refine unknown macromolecular structures based on NMR or X-ray data.[1] In terms of gene technology, it can be used to predict 3D protein structures from a primary sequence via homology modeling methods.[2] In addition, computational chemistry can help in the design and optimization of new drugs and pharmaceuticals. Computational Chemistry typically offers several advantages over the usual approach of performing and analyzing experiments. These include low cost, high safety, and reasonable accuracy. They also provide significantly more information at atomic level.

Clearly, the computational study of biological systems has shown rapid progress in recent years. Meanwhile, significant improvements in the general methodology have been made in order to allow greater accuracy and flexibility for a wild variety of utilization. Examples include the availability of improving force fields, the accurate treatment of long-range atom–atom interactions, and new optimization algorithms.[3-5] One of the most popular approaches to testify the force field improvement is calculating the energy changes like conformational free energy differences and barriers, and comparing it with experimental results.

Generally speaking, we can divide the common computational methods into two classes: Classical Mechanics (CM) and Quantum Mechanics (QM). QM provides the most accurate approach. It can calculate almost every molecular property of interest including structure, conformational energies, dipole moments, ionization potentials, electron affinities, electron densities, etc. However, the computational expense of QM methods has limited applications to relatively small molecules. All CM approaches derive from the Born-Oppenheimer approximation - stating that the Schrodinger Equation can be separated into electron and nuclei motion, which can be treated independently. Hence, the major attention is focused on the motion of nuclei under the assumption that the electrons have already found their optimal distribution. The energy of the whole system can then be considered as a function of just the atomic nuclear

coordinates. It is much more efficient than QM calculation. As we are more interested in biological systems, CM becomes the only general approach available. In CM approaches one uses approximate empirical functions and parameters to provide a description for the energy of a particular collection of atoms.

Furthermore, using various kinds of mathematical algorithms, CM is usually applied in three main approaches, namely, molecular mechanics (MM) calculations, Monte Carlo (MC) simulation, and molecular dynamics (MD) simulation.[6] MM focuses on energy minimization based on Newton's laws. Minimization calculation cannot detect all conformational possibilities. In another word, it lacks the ability to recognize local minimum and global minimum if a large energy barrier exists. MC is based on the behavior of stochastic process and probability distributions provided by statistical mechanics. The drawbacks of this method are related to the fact that it cannot provide any dynamical information, and the implementation for polymer chains is somewhat complicated. MD is the most popular and widely used method, especially for systems like biomolecules, polymers, metals, and non-mental materials. The application of Newton's laws of motion, and other restrictions which will be introduced in following sections, allows one to follow the movement and behavior of large collections of particles.

Hence, MM emphasizes minimizing the potential energy in a system; MC relies on statistical probabilities; while MD focuses on modeling molecular motions. We would like to investigate and improve molecular models using MD, considering that MD is the most prevalent and suitable for biological systems.

1.2 Protein Folding Mechanism

One of the major aims for biological simulation is to correctly predict the 3D structure of proteins, and thereby discover the driving forces behind protein folding.[3] This type of approach typically assumes the final conformation is unique and corresponds to lowest free energy arrangement of the atoms for a given environment. This can only be achieved computationally if one has accurate descriptions of all the interactions involved in the system. Here, we briefly discuss several different kinds of non-covalent interactions considered as providing the most important contributions to protein folding.

1.2.1 Van Der Waals Interaction

The van der Waals interactions include both attractive forces and repulsive forces. The attractive forces arise from dipole interactions due to fluctuations in the electron densities between all atom pairs. The repulsion is a consequence of the electron-electron repulsion that occurs as two clouds of electrons begin to overlap. Although van der Waals interactions are very weak in magnitude compared to other forces contributing to protein folding, it is often observed that the large number of such interactions between tightly packed groups in the interior of a large protein that play an essential role in protein folding.[7,8]

1.2.2 Electrostatic Interaction

At neutral pH some amino acid side chains contain positively charged groups (lysine, arginine, and histidine), while some contain negatively charged groups (aspartate and glutamate). In addition, the protein terminal residues often exist in the ionized state ($-NH_3^+$, $-COO^-$) in aqueous solution. In most cases, charged residues are located on the protein surface, interacting with water to help stabilize native protein structure. Meanwhile, the presence of salts in the solution can also influence electrostatic interactions.

Usually, electrostatic interactions are affected by pH and/or ionic strength. The pH determines the total charge on the protein, while salt can shield charge interactions within the protein or between proteins. They influence the stability of proteins in different ways. For instance, one can usually denature a protein by varying the solution pH. The pH determines the protein net charge. On adding acid or base into the solution the net charge on the protein will become more positive or negative, respectively. The resulting repulsion between side chain groups will then increase and eventually destabilize the native protein structure.[9-11] Not all salt effects contribute to destabilizing protein folding; ion pairing sometimes could stabilize it as well. Usually it is assumed that salts do not affect the net charge on the protein. There have been many studies into the role of ion pairing and possible contributions to protein stability,[12-14] but it does not appear to be a dominant force in protein folding due to the small number of ion pairs in proteins.

1.2.3 Hydrogen Bonds

Proteins contain many hydrogen bond donors and acceptors in both the peptide backbone and the amino acid side chains. Consequently, hydrogen bonds can be generally formed

internally within protein itself or between protein chains, and also externally with other surrounding molecules. Atoms from the protein backbone tend to form hydrogen bonds with one another and are central components for secondary structure - between the C=O and N-H groups in α-helices and β-sheets, for example. Amino acid side chains on the protein surface can form hydrogen bonds with water solvent molecules. Due to the large amount of hydrogen bonds formed on folding, the total hydrogen bonding effects are presumed to have a strong impact on the folding pathway. However, many studies have indicated that hydrogen bonding is less important than hydrophobic interaction,[15-17] and therefore it is generally not considered the dominant folding force.[18,19]

1.2.4 Hydrophobic Interaction

Proteins contain amino acids with either hydrophilic or hydrophobic side chains. How these different side chains interact with an aqueous solution can play a major role in protein folding. As mentioned above, polar side chains of amino acids would like to stay at protein surface where they remain solvated. However, non-polar side chains, such as leucine, isoleucine, phenylalanine, tryptophan, prefer to be buried in the interior of globular proteins which represents a relatively non-polar environment. The main reason for this is that the non-polar side chains are composed from hydrocarbon type groups. They cannot form hydrogen bonds with water and, in an effort to maintain as many water-water hydrogen bonds as possible, non-polar residues tend to combine together and form buried hydrophobic regions. This process minimizes the interaction of hydrophobic residues with water and is therefore highly energetically favorable. The formation of such hydrophobic cores is therefore entropy driven. It is usually considered the major driving force of protein folding.[9,20,21]

1.3 Molecular Dynamics (MD) Simulation

1.3.1 Introduction

Molecular dynamics simulation is widely used in chemistry, physics, and materials science. It provides information concerning the detailed physical movements of atoms and molecules. This is an important technique for the study of the natural time evolution of a system and can be used to predict the static and dynamic properties at equilibrium. Dynamical information can be obtained by continuously solving the equations of motion for a series of finite

steps in time. Depending on the demand for accuracy, the equations of motion to be solved could be simply the classical equations of motion (Newton's), the stochastic equations of motion (Langevin's), the Brownian equations of motion, or even a combination of quantum and classical mechanics.

1.3.2 The Statistical Mechanics Basis of Molecular Dynamics

Using classical Hamiltonian mechanics the Hamiltonian, H, is equal to the total energy of a system as a function of both coordinates (r) and momenta (p). The potential energy U(r) is independent of time and velocity, while the kinetic energy K(p) is the sum of kinetic energy of all particles in the system,

$$H = \mathrm{H}(\mathrm{r},\mathrm{p}) = \mathrm{K}(\mathrm{p}) + \mathrm{U}(\mathrm{r}) = \sum_{\mathrm{i}} \frac{\mathrm{p_i^2}}{2\mathrm{m_i}} + \mathrm{U}(\mathrm{r}) \tag{1.1}$$

Statistical mechanics (SM) provides a link between properties of collections of particles, as provided by a computer simulation, and the resulting thermodynamic properties of the system. To obtain thermodynamic averages we need the probability distribution function, ρ(r, p), shown as below.

$$\rho(\mathrm{r},\mathrm{p}) = \frac{\exp[-\mathrm{H}(\mathrm{r},\mathrm{p})]/\mathrm{k_B T}}{\mathrm{Z}} \tag{1.2}$$

where Z is the partition function over a canonical ensemble (*N*, *V*, *T*), and k_B is the Boltzmann factor. If we know the probability distribution function, we can calculate average values of properties *A*(r, p), like position, energy, etc.

$$\langle \mathrm{A}(\mathrm{r},\mathrm{p}) \rangle_{\mathrm{Z}} = \int_{\mathrm{V}} \mathrm{dr} \int_{-\infty}^{\infty} \mathrm{dp}\, \rho(\mathrm{r},\mathrm{p}) \mathrm{A}(\mathrm{r},\mathrm{p}) \tag{1.3}$$

However, in order to calculate these averages it is necessary to simultaneously know the Boltzmann probability [Eq. (1.2)] for every state defined by a specific set of r and p, which is an extremely difficult computational task.

An alternative strategy is to follow the motion of a collection of particles as a function of time. Then, the averages are calculated only over those points that were passed by during the path instead of averaging over all possibility. For example, starting from a point r(0), p(0), the

procedure yields a trajectory describing the system at time t as r(*t*), p(*t*) within a simulation of duration time τ. Such kinds of averages are called "dynamic averages" and are calculated as follows.

$$\langle A(r,p)\rangle_\tau = \frac{1}{\tau}\int_0^\tau A\big(r(t),p(t)\big)\,dt \tag{1.4}$$

Hence, thermodynamic properties can be obtained from either a weighted average over all members of an ensemble, or as dynamic average over time. Time averaging is generally easier to perform compared to the ensemble average approach.

The two averaging strategies are therefore connected. The assumption is that, for an infinitely long trajectory, the thermodynamic ensemble average and the dynamic average become equivalent to each other. This is also known as the ergodic hypothesis.

$$\lim_{\tau\to\infty}\langle A(r,p)\rangle_\tau = \langle A(r,p)\rangle_Z \tag{1.5}$$

The ergodic hypothesis tells us that when the trajectory becomes long enough, the two averages become identical. Therefore, we can use molecular dynamic simulations to calculate thermodynamic averages of molecular systems.

1.3.3 Newtonian Molecular Dynamics

Even though the molecular behavior is best described by quantum mechanics, the computation is very expensive for large systems. Therefore, classical mechanics methods are often used for biomolecular systems. In our studies we will use Newtonian equations of motion (primarily Newton's second law of motion) so that,

$$-\nabla_i U(r) = F_i = m_i a_i = m_i \ddot{r}_i \tag{1.6}$$

where F_i is the force on particle i, m_i is the mass of particle i, a_i is its acceleration, and $\ddot{r}_i$ is the second derivative of the particle position r with respect to time. The force F_i is determined by the gradient of the potential energy function, U(r), which is a function of all the atomic coordinates r. In the MD method the net force on a molecule, arising from all the other molecules in the system, is evaluated for an initial atomic arrangement. Each molecule in the system is then

moved using Newton's law for a short time interval. The forces are then recalculated and the whole process repeated to provide with the dynamic properties of the system.

Solving Newton's equations of motion requires a numerical procedure for integrating the above set of differential equations. Typically, the molecular coordinates and velocities at a time $t + \Delta t$ are obtained from the molecular coordinates and velocities at an earlier time t. The equations are solved on a step-by-step basis. The choice of time interval Δt depends on the properties of the molecular system simulated, andΔt must be significantly smaller than the fastest characteristic time of the system to be studied.

A good starting point for understanding such finite-difference methods is the Taylor expansion about time t to obtain (predict) the position at time $t + \Delta t$,

$$\mathrm{r}(\mathrm{t} + \Delta\mathrm{t}) = \mathrm{r}(\mathrm{t}) + \mathrm{v}(\mathrm{t})\Delta\mathrm{t} + \frac{1}{2}\mathrm{a}(\mathrm{t})\Delta\mathrm{t}^2 + \cdots \tag{1.7}$$

where v(t) is the velocity vector and a(t) is the acceleration. Because the integration proceeds in a stepwise fashion, and recalling Equation 1.6, it is convenient to rewrite the above expansion in a discrete form. Using $\mathrm{r_n}$ to indicate the position at step n (at time t) and $\mathrm{r_{n+1}}$ to indicate the position at the next step, n+1 (at time t+Δt), Equation 1.7 can be written as

$$\mathrm{r_{n+1}} = \mathrm{r_n} + \mathrm{v_n}\Delta\mathrm{t} + \frac{1}{2}\left(\frac{\mathrm{F_n}}{\mathrm{m}}\right)\Delta\mathrm{t}^2 + \mathrm{O}(\Delta\mathrm{t}^3) \tag{1.8}$$

where $\mathrm{O}(\Delta\mathrm{t}^3)$ is the terms of order $\Delta\mathrm{t}^3$ or smaller. With this information the velocity $\mathrm{v_{n+1}}$ at time $n + 1$ can be crudely estimated, for example, as

$$\mathrm{v_{n+1}} = (\mathrm{r_{n+1}} - \mathrm{r_n})/\Delta\mathrm{t} \tag{1.9}$$

Together, Equation 1.8 and 1.9 form an integration algorithm. Given the position $\mathrm{r_n}$, the velocity $\mathrm{v_n}$, and the force $\mathrm{F_n}$ at step n, these equations allow one to calculate the position $\mathrm{r_{n+1}}$ and velocity $\mathrm{v_{n+1}}$ at step n+1. There are many algorithms which provide variations on Equation 1.8.[22]

1.4 Force Field Development for Biological Systems

1.4.1 Force Field and Potential Energy Function

A force field is a collection of equations and parameters for use in determining the potential energy of a system. The parameters might come from fitting against experimental data or quantum mechanics calculations. The key to the success of studying chemical systems by computer simulation is the quality of mathematical models which are used to calculate the energy of the system as a function of the coordinates.

Generally speaking, there are two kinds of force fields in common use depending on the intended application. Class I force fields express interactions in a relatively rigorous way since they aim to optimize small molecules, especially organic molecules. Examples of this class include MM3,[23,24] consistent force field (CFF),[25] COMPASS,[26] and Merck Molecular Force Field (MMFF).[27] They focus on optimizing geometries and energetics with high accuracy. Two other force fields with a similar aim are the UFF[28] and DREIDING[29] approaches, which are specifically developed to treat inorganic compounds. The Class II force fields include a variety of force fields used for bio-macromolecular systems. The most popular packages currently available include CHARMM, AMBER, GROMOS, OPLS, etc.[3,30-39] Most of them were developed to perform simulations on biological systems such as proteins,[40] nucleic acid bases,[41] and carbohydrates.[4]

Clearly, the wide variety for force fields available indicates that one should carefully choose which is the most appropriate for a particular application. However, the general purpose of a force field (class I or II) is to use a variety of energy functions to accurately describe a range of molecular properties and inter molecular interactions. The energy functions usually contain both bonded and non-bonded interactions: bonded interactions include stretching (bond), bending (angle), and rotation (proper and improper dihedrals); non-bonded interactions include electrostatic interactions (long-range forces) and van der Waals potentials.

The energy equations used for the study of biological systems often involve a series of terms. An example of the terms typically used is provide below and will be used in later chapters. The energy is divided in to contributions from the various degrees of freedom in the molecules of interest such that,

$$\sum_{bonds} \frac{1}{2} k_r (r - r_0)^2 \tag{1.10}$$

$$\sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \tag{1.11}$$

$$\sum_{improper} \frac{1}{2} k_\xi \left(\xi_{jikl} - \xi_0\right)^2 \tag{1.12}$$

$$\sum_{torsions} k_\phi [1 + \cos(n\phi - \phi_s)] \tag{1.13}$$

$$\sum_{elec} \frac{q_i q_j}{r_{ij}} \tag{1.14}$$

$$\sum_{LJ} 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \tag{1.15}$$

The first four equations correspond to the bonded interactions (bond, angle, improper dihedral and proper dihedral). All the bond-stretching, angle-bending and improper dihedral terms are treated harmonically, based on Hooke's Law, which effectively keeps the bonds and angles near their equilibrium values at room temperature. Bond and angle parameters include r_0, θ_0, and ξ_0, the equilibrium bond length, equilibrium angle and equilibrium improper dihedral, respectively. k_r, k_θ and k_ξ are the force constants associated with the bond, angle and improper dihedral terms, respectively. Proper dihedral, also called torsion angle, represent the variation in energy that can occur for rotation about a bond. The dihedral term includes parameters for the force constant, k_ϕ; the periodicity, n; and the phase, ϕ_s.

The last two equations describe the non-bonded interactions (electrostatic and van der Waals interactions). These terms represent the most important energy contributions for computational studies of biological systems, due to the strong environmental influence and the large number of non-bond interactions. Here, the electrostatic interaction depends on the distance

between atoms i and j, r_{ij}, and the partial atomic charges, q_i for each site (usually atom based) on the molecule. The van der Waals interaction is often modeled using a Lennard-Jones (LJ) 6–12 potential as indicated above. The strong distance dependence of the repulsion is indicated by the $1/r^{12}$ power term, while the London's dispersion interaction or instantaneous induced dipole–induced dipole interaction is provided by the $1/r^6$ term, with a negative sign indicating a favorable contribution.

1.4.2 Computational Efficiency versus Accuracy

As mentioned previously, there is always a conflict between accuracy and computational efficiency. There are several factors which affect the accuracy and efficiency of a simulation which we will discuss briefly. All-atom models, like Charmm22,[31] Amber99SB,[42] OPLSAA,[43] include every atom contained in the molecule explicitly; thereby requiring a relatively longer time for calculation. To accelerate the calculations it is common to use united-atom models instead all-atom models. The main difference being that in united-atom models the non-polar aliphatic hydrogens are not explicitly represented, but rare treated as part of the non-hydrogen heavy atom to which they are covalently attached. Polar hydrogens, however, are important for representing hydrogen bonding interactions and are therefore treated explicitly. Several force fields, such as Gromos[44] and OPLS[40], are united-atom models which attempt to provide precise results but more efficiently.

Another type of model which aims at significant improvements in computational efficiency are general known as is coarse grain models. An example of this type of approach is the Martini model.[45] Here, the peptide chain is replaced with groups of atoms, or beads, which are treated as one site and therefore greatly simplify the calculations. Furthermore, multiple solvent molecules are replaced by a single solvent bead. These simplifications greatly speed up the calculations, but introduce approximations into the description of the interactions between molecules. They are often applied to study very large biological systems such as proteins in membrane. Coarse grain models will not be considered here.

Most biological systems of interest include a protein surrounded by solvent molecules. When all water molecules are included in a simulation it is often referred to as an explicit solvent simulation. This is the most accurate solvent representation, but also the most expensive. Attempts to remove the solvent molecules, while attempting to model the effects of solvation, in

an effort to improve efficiency have been investigated. These are known as implicit solvent models and have been widely used to study small peptide folding.[46] They generally treat the solvent using continuum electrostatic approaches, e.g. Poisson-Boltzmann (PB) and Generalized Born (GB) models. However, many specific solvent effects – such as the hydrophobic effect, the viscosity of solvent and hydrogen bonding with solvent – are difficult to capture accurately. Implicit solvent approaches also perform poorly for larger proteins.[47-49]

Finally, it is well known that polarization in the condensed phase can affect the intermolecular interactions. All the force fields mentioned above are non-polarizable force fields which basically ignore any induced polarization effects. In an effort to improve force field accuracy, several polarizable force fields have arisen attempting to compensate for this defect.[5,50] The introduction of explicit polarization into a model is relatively easy. However, it requires new parameters and has the disadvantage of requiring computationally expensive iterative approaches to solve for the polarization forces. Hence, we will not consider these force fields in any detail.

1.4.3 Water Models

Essentially all biomolecular simulations are conducted in the condensed phase, which almost always involves a solution of water molecules. Thus, there exist several different kinds of non-bonded interactions - the solvent–solvent (e.g., water–water), solvent–solute (e.g., water–protein), and solute–solute (e.g., protein-protein or protein intramolecular) interactions.[51,52] Hence, it is very important to use a reliable water model for biomolecular simulations to ensure a reasonable balance between these interactions. The most popular water models in current use include the TIP3P,[53] SPC[54] and SPC/E[55] models. The SPC/E water model is known to yield better pure solvent properties than the TIP3P and SPC models. Hence, most of the simulations and studies provided here have been performed using the SPC/E model.

1.4.4 Future Directions

There are many attempts to provide accurate force fields for the simulation of peptides and proteins. Many successful simulations of proteins have been performed. However, the force fields still possess significant weaknesses and require constant improvement. In particular, it appears that one can often accurately reproduce biologically relevant peptide and protein structures, but many thermodynamic properties don't match very well with experimental data.[56-58] Thus, future efforts will continue to optimize and refine empirical force fields.[59-61]

Furthermore, the need to pay more attention to the associated thermodynamics while improving a force field, has led to the general approach presented here. This approach is based on the Kirkwood-Buff Theory of solutions.

1.5 Kirkwood-Buff (KB) Theory

The Kirkwood–Buff (KB) theory of solutions was proposed by Kirkwood and Buff in 1951.[62] The theory is totally general and can be applied to all kinds of solutions over the entire range of compositions. It is an exact theory with no approximations, which makes it more valid than other theories.[63] The theory provides a direct relationship between molecular distributions at the atomic level and bulk thermodynamic properties, e.g. partial molar volumes, chemical potentials and compressibility. Ben-Naim later developed the inversion procedure of KB theory,[64] providing information about the affinity between a pair of species in the solution mixture from experimental thermodynamic properties. Symbolically, the original theory may be written as $\{g_{ij}\} \rightarrow$ {Thermodynamic quantities}; while the inversion theory may be written as {Thermodynamic quantities} $\rightarrow \{G_{ij}\}$ where the quantities G_{ij} represent integrals over the molecular distributions. The theory has become more and more popular and widely applied by many scientists to a variety of processes. These include Smith,[65-87] Marcus,[88] Ruckenstein,[89-98] Shimizu,[99-103] Hall,[104] Zielkiewicz,[105] Lepori,[106,107] O'Connell,[108,109] etc. In addition, many chemists and physicists are continually developing KB theory and applying it to study solution mixtures.[66,71,72,77-105,110-126]

The relative distribution of particles in a system can be expressed using radial distribution functions. A radial distribution function (rdf), $g(r)$, provides the probability of finding a particle at a distance r around a central particle. It describes how the solution density varies as a function of the distance. In an N particle system, the probability function for finding particle 1 at distance dr_1 and particle 2 at distance dr_2 can be expressed as following using Boltzmann distribution:[127-129]

$$P(r_1, r_2) = \frac{\int\int\cdots\int e^{-\beta U_N}\, dr_3 dr_4 \cdots dr_N}{Z_N} \tag{1.16}$$

where $\beta = 1/kT$, U_N is the N-particle potential energy, and Z_N is known as the configurational integral. Consequently, the probability of finding any atom at distance dr_1, and any atom at distance dr_2, can be written as

$$\rho(r_1, r_2) = \frac{N!}{(N-2)!} P(r_1, r_2)$$

(1.17)

The probability of finding a particle anywhere in the system would be generally expressed as

$$\frac{1}{V}\int \rho(r_1) dr_1 = \rho = \frac{N}{V}$$

(1.18)

Therefore, $g(r)$ can be introduced as

$$\rho(r_1, r_2) = \rho^2 g(r_1, r_2)$$

(1.19)

which is provided by equations 1.16, 1.17, 1.18, 1.19 as,

$$g(r_1, r_2) = \frac{V^2 N!}{N^2 (N-2)!} \frac{\int\int \cdots \int e^{-\beta U_N} dr_3 dr_4 \cdots dr_N}{Z_N}$$

(1.20)

and measures the deviation from random, or correlation, but particles of different types. Figure 1.2 shows a typical radial distribution function. It starts from 0 at short distances due to the strong repulsion between two particles. Then it typically displays a series of fluctuations around $g(r) = 1$, which are generally known as solvation shells. The first peak, which is also the largest one, indicates that one is most likely to find a particle at this distance, compared to other distances, with respect to the bulk solution distribution. As the distance r increases, the distribution of components approaches unity, which indicates a random bulk solution distribution. Radial distribution functions can also obtained from experiment using X-ray diffraction studies of solutions. But this is only useful for small molecules.

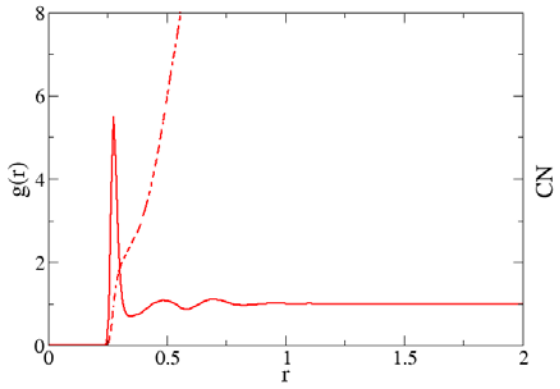
The integration of a radial distribution function between two different species i and j provides a Coordination Number,

$$CN(i,j) = \rho_j \int_0^R g_{ij}(r) 4\pi r^2 dr$$

(1.21)

to a distance R from the central molecule. An example of an associated coordination number is also provided in Figure 1.2.

Figure 1.2 An example of a typical radial distribution function *g*(*r*) (solid line) and coordination number CN (dashed line) as a function of distance *r* (nm).



The radial distribution function provides insight into the liquid structure. The corresponding integrals over g(r), also called KB integrals (KBIs), are useful to express thermodynamic properties of solution mixture, such as compressibility, partial molar volumes and derivatives of the chemical potentials.[82,83,85-87] Hence, combinations of KB integrals provide a link between thermodynamic properties and molecular distribution functions for multi-component systems. The KB integrals are expressed by the following equation:

$$G_{ij} = 4\pi \int_0^{\infty} \left[g_{ij}^{\mu VT}(r) - 1 \right] r^2 dr \tag{1.22}$$

where G_{ij} is KB integral between species i and j, $g_{ij}^{\mu VT}$ is the corresponding radial distribution function in the μVT ensemble, r is the corresponding center of mass - to - center of mass distance. Thus, the theory may be used to compute the thermodynamic quantities of the pair correlation function.

An excess coordination numbers can be defined from the KBIs according to

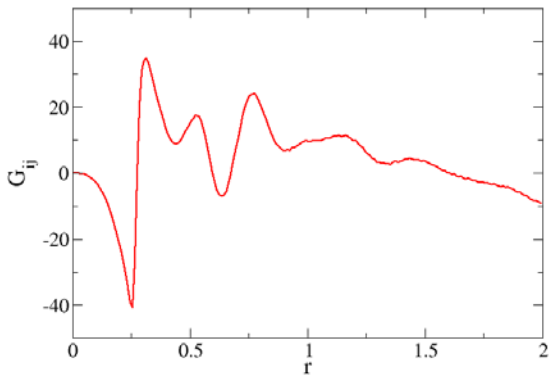
$$N_{ij} = \rho_j G_{ij} \tag{1.23}$$

where ρ_j is the number density (molar concentration) of species j.

$$\rho_j = N_j / V \tag{1.24}$$

A value of N_{ij} greater than zero indicates an excess of species j in the vicinity of species i over a random distribution, while a negative value corresponds to a depletion of species j surrounding i. In other words, a positive N_{ij} can be interpreted as favorable (attractive) interactions between species i and j, and a negative N_{ij} is related to unfavorable (repulsive) interactions. Generic examples of KB integrals G_{ij} and excess coordination numbers N_{ij} are illustrated in Figure 1.3 and 1.4. They provide a sensitive test of the relative distribution of the different species in solution.[87]

Figure 1.3 An example of a typical KB integral G_{ij} as a function of integration distance r (nm) obtained from computer simulation.



The Figure 1.4 shows excess coordination number obtained for n-propanol and water mixtures (see Chapter 2 for details). The dashed line indicates water-water pairs, while the solid line shows n-propanol-water pairs. The KBIs vary with composition and provide quantitative information on the relative distributions in solution. Clearly, at a solute mole fraction of 0.2 we can observe apparent aggregation among water molecules and exclusion of n-propanol molecules surrounding water. These types of features are important and interesting. A reasonable force field for the simulation of solution mixtures should be able to reproduce such trends.

In the KB theory of solutions thermodynamic properties of a solution mixture can be derived from radial distribution functions, and vice versa. Figure 1.5 illustrates the role of KB theory as a bridge between these two. KB integrals, G_{ij}, are defined as integrals over radial distribution functions between species i and j.[73]

Figure 1.4 An example of excess coordination number N_{ij} as a function of solute mole fraction obtained from experimental data. The example is for binary mixtures of n-propanol and water (see Chapter 2 for details).

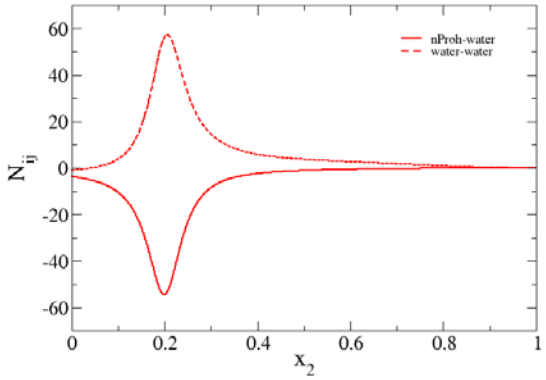
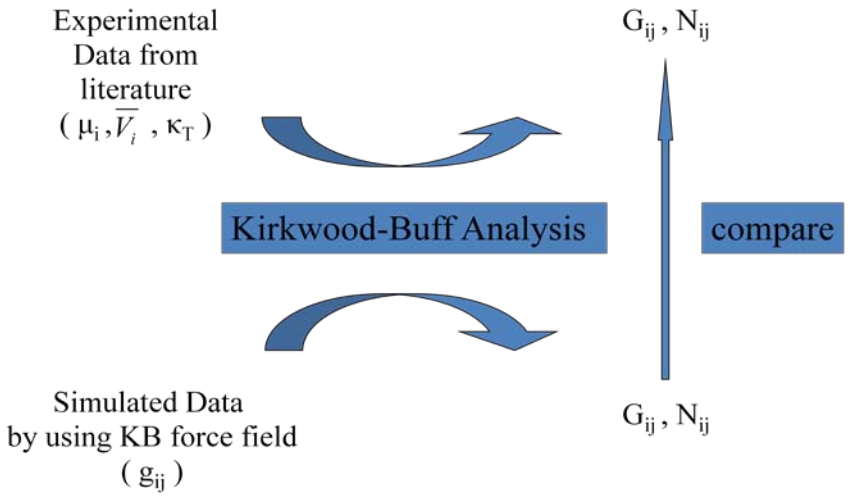


Figure 1.5 A schematic illustration of a typical KB analysis relating both experimental and simulation data.



Hence, KB integrals can be determined either from experimental or simulated data. For a solution mixtures with water and solute at constant pressure (p) and temperature (T), the chemical potentials (μ_i), partial molar volumes ($\overline{V}_i$), and isothermal compressibilities (κ_T) can be obtained experimentally. Then the experimental data can be used to determine KB integrals.[101]

KB theory can also be applied to biomolecular systems, as well as cosolvent systems to analyze the free energy of molecular binding and characterize the preferential interactions and other thermodynamic properties which we will discuss details in following chapters. In a system of a biomolecule (2) and cosolvent (3) with primary solvent of water (1), the preferential binding parameters can be obtained from equilibrium dialysis experiments and also expressed using KB integrals.[79]

1.6 Kirkwood-Buff Derived Force Field

The key to accurate biomolecular simulation is to develop high quality force fields for proteins. It is noticed that current available force fields tend to over stabilize secondary structure;[130] some are too helical heavy and some may be beta-sheet biased. Significant work to develop and improve force fields has been performed, including Brooks,[131] Jorgensen,[132] Berendsen,[133] Levitt,[134] van Gunsteren[135] and others,[136-138] but current available force fields can still be improved. In particular, they struggle to reasonably reproduce some physical properties.[82,83,86,139]

A possible avenue for improvement involved the solvation interaction. It is believed that part of the force field inaccuracies can be traced to the approximate treatment of polarization effects using effective partial atomic charges, which leads to an imbalance between the solute-solute, solvent-solvent interactions due to an underestimation of the solute-solvent interactions.[83,139-142] Most effective charge distributions for molecules are provided by gas phase QM calculations, rather than the more appropriate (but expensive) condensed phase calculations. Gas phase calculations only take into account the permanent mutipole moments with no solvation interaction involved. Unfortunately, the ignored solvation effect can lead to significant changes to charge distribution which should not be ignored. Hence, most empirical force fields provide only an approximate representation of the molecular polarity in condensed phases. This severely limits the reliability and predictability of molecular properties in biological systems. A simple and highly accurate description of the charge distribution in solution is therefore required.

As mentioned previously, one of the possible developments in force field design is the use of explicit polarization approach to achieve more accurate results. In principle, this should provide more realistic and accurate results than non-polarizable force fields. However, the additional computational cost expensive and difficultly finding a unique method to treat pair-

wise polarizable interactions has been problematic.[50,143-148] Thus, non-polarizable force fields are still the most popular and widely used approach. In contrast, non-polarizable force fields developers have tried to simply rescale charge distribution in order to distinguish between the gas and solution phases, but it has been pointed suggested that the electronic rearrangements occurring in the solvation process is far more complicated than provided by simple scaling from the gas phase. Therefore, we will introduce an alternative way to improve the accuracy of simple non-polarizable force fields by taking advantage of Kirkwood-Buff Theory. This approach will mainly focus on correctly reproducing the effective charge distributions for molecules in condensed phase by reproducing experimental values of the KBIs.

As noted previously,[65,82,83,86,87,110] KB theory is a powerful tool to evaluate the ability of a force field to correctly represent relative molecular distribution in solutions. KB theory is an exact theory of solution mixtures and valid for the analysis of both experimental and theoretical solvation quantities with no limitations to the size or character of molecules. The quality of a force field used for simulation can be easily determined by comparing KB integrals derived from simulated data to those extracted from the experimental data. In addition, the KB integrals are more sensitive to the parameters from force fields than many other thermodynamic properties,[82,83,85-87] which provides a solid basis for judging accuracy of a particular force field. For instance, the KB integrals are directly related to the molecular affinity information which is a consequence of the interactions among the atoms.

Unfortunately, many existing force fields perform poorly in their ability to reproduce the experimental KB integrals.[110] This indicates that currently used force fields do not correctly reproduce the solution distributions,[110] and this can lead to inaccurate simulation results. Therefore, it is necessary to develop an improved force field which can truly represent the correct molecular distributions in a solution mixture, and thereby maintain a reasonable balance between solute-solute interactions and solute-solvent interactions (solvation). This is the aim of the Kirkwood-Buff derived force field (KBFF) approach. During the past several years the Smith group has been developing Kirkwood-Buff derived force fields as a central aspect of their work. The only major difference to other similar biomolecular force fields is the origin of the effective charge distributions. Other parameters are similar to most common force fields. Molecular geometries are obtained from experimental data on crystal structures. Bond parameters are taken from the GROMOS96.[149] The general non-bonded form of the KB force field contains a

Lennard–Jones (LJ) 6-12 potential plus Coulomb interaction. The molecular charge is explored thoroughly during the parameterization process, while the van der Waals parameters for hydrocarbons were taken from elsewhere.[149] It has been shown that simulation results from the KBFF models perform fairly well and can be even better than other common force fields with similar computational cost.[3,40,150-152]

Below is a list of recent publications regarding our force fields development using the Kirkwood-Buff theory of solutions as a guide:

1. Weerasinghe and Smith, JCP, v118, 5910, 2003.	Urea
2. Weerasinghe and Smith, JCP, v118, 10663, 2003.	Acetone
3. Weerasinghe and Smith, JCP, v119, 11342, 2003.	NaCl
4. Weerasinghe and Smith, JCP, v121, 2180, 2004.	GdmCl
5. Weerasinghe and Smith, JPCB, v109, 15080, 2005.	Methanol
6. Kang and Smith, JCC, v27, 1477, 2006.	NMA
7. Gee and Smith, thesis, 2010.	Small amino acid
8. Bentenitis, Cox and Smith, JPCB, v113, 12306, 2009.	Thiols, sulfides, disulfides
9. Ploetz and Smith, PCCP v13,18154 2011.	Aromatic amino acid
10. Gee, Smith et al, JCTC v7, 1369, 2011.	Alkali halides
11. Ploetz and Smith, Fluid Phase Equilibria v290 43 2010.	Pyrrole and indole
12. Dai, Weerasinghe and Smith, to be published.	Carboxylic acid, amines

In this thesis we continue this research to provide a full force field capable of simulations of peptides and proteins in a variety of solutions. The outline of the thesis is as follows:

In Chapter 2 we will present details concerning the transferability of the force field models between small organic molecules. In Chapter 3 we will then apply our force field for the study of three component (ternary) solution mixtures involving water with simple peptide model (N-methylacetamide) and a cosolvent (urea).

The essential problem for computational studies of biological systems is accurate simulations of protein folding and/or aggregation. Successful simulations of this kind heavily depend on the quality of the force fields used.[42,151] In order to achieve our goal to represent the correct conformational distributions of peptides and proteins in solution, it is important to balance the protein backbone parameters, especially the backbone phi/psi potentials. Existing force fields, such as CHARMM,[150] OPLS,[40] AMBER[3] and GROMOS[149] often exhibit some bias

or limitation in reproducing the appropriate secondary structure distributions. For example, it has been reported that high propensities towards helices were observed,[42,43,151] due to incorrect force field parameters. Hence, we require accurate protein backbone Φ/Ψ potentials. In Chapter 4 we will present our approach to develop KBFF parameters for biological systems, specifically the torsional conformations of the protein backbone.

1.7 Summary and Outline

Molecular dynamic simulations have played a key role in the study of biological systems and provide information at the atomic level which is not available experimentally. Kirkwood-Buff theory can be used to interpret experimental and computational data and to provide a bridge between them. Here, we use KB theory and computer simulations for a variety of applications.

In Chapter 2 a Kirkwood-Buff derived force field for a series of alcohols in water is developed. The alcohols include methanol, ethanol, n-propanol, iso-propanol, n-butanol, tert-butanol and n-octanol. KB integrals are calculated and compared with the available experimental data. Also a recent, more general, fluctuation theory is applied to investigate systems like methanol, ethanol and n-prpanol and iso-propanol, to further validate the good performance of the current force fields.

In Chapter 3 a three component solution is investigated by Kirkwood-Buff theory and computer simulation. A system of NMA (0, 1, 2, 4 m) in urea (0, 2, 4, 6, 8, 10 m) solution is studied. NMA is considered as a simple model for a peptide bond in biological systems, while urea is chosen as a protein denaturant. KB theory is used to calculate the transfer free energy and preferential interactions for further understanding the mechanism of protein denaturation by urea at the atomic level.

In Chapter 4 we developed and test a Kirkwood-Buff derived force field for peptides and proteins. To achieve our goal of a complete Kirkwood-Buff derived force field for proteins, we provide improved parameters for the protein backbone torsional dihedrals. The proposed φ/ψ backbone potentials are then tested using a variety of small peptides and proteins.

1.8 References

(1) Smith, S. J.; Sutcliffe, B. T. *The Development of Computational Chemistry in the United Kingdom*; VCH **1997**.

(2) Attwood, T. K.; Gisel, A.; Eriksson, N. E.; Bongcam-Rudloff, E. In *Bioinformatics - Trends and Methodologies.*; InTech: **2011**.

(3) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179.

(4) Damm, W.; Frontera, A.; TiradoRives, J.; Jorgensen, W. L. *Journal of Computational Chemistry* **1997**, *18*, 1955.

(5) Gao, J. L.; Habibollazadeh, D.; Shao, L. *Journal of Physical Chemistry* **1995**, *99*, 16460.

(6) Garcia Valcarcel, A. *Med Esp* **1952**, *27*, 283.

(7) Ptitsyn, O. B.; Finkelstein, A. V. *International Journal of Quantum Chemistry* **1979**, *16*, 407.

(8) Englander, S. W.; Mayne, L.; Krishna, M. M. G. *Quarterly Reviews of Biophysics* **2007**, *40*, 287.

(9) Kauzmann, W. *in The Mechanism of Enzyme Action*; Johns Hopkins Press: Baltimore, MD, **1954**.

(10) Linderstrom-Lang, K. U. In *C. R. Trau. Lab. Carlsberg* **1924**, *15*, 70.

(11) Tanford, C. In *Physical Chemistry of Macromolecules, Chapter 7*; Wiley: New York, **1961**.

(12) Eyring, H., Stearn, A. E. *Chem. Rev.* **1939**, *24*, 253.

(13) Green, A. A.; Cohn, E. J.; Blanchard, M. H. *J. Biol. Chem.* **1935**, *109*, 631.

(14) Mirsky, A. E.; Pauling, L. *Proc Natl Acad Sci U S A* **1936**, *22*, 439.

(15) Tanford, C. *Adv Protein Chem* **1968**, *23*, 121.

(16) von Hippel, P. H.; Schleich, T. In *Structure and Stability of Biological Macromolecules. Vol. II*; Timasheff, S., Fasman, G., Eds.; Dekker: New York, **1969**.

(17) Herskovi, T.; Gadegbek, B.; Jaillet, H. *Journal of Biological Chemistry* **1970**, *245*, 2588.

(18) Singer, S. J. *Adv. Protein Chem.* **1963**, *17*, 1.

(19) Edelhoch, H.; Osborne, J. C., Jr. *Adv Protein Chem* **1976**, *30*, 183.

(20) Linderstrom-Lang, K. U. In *Lune Medical Lectures, Vol. 6* Stanford University Press: Stanford, CA, **1952**.

(21) Lumry, R.; Eyring, H. *Journal of Physical Chemistry* **1954**, *58*, 110.

(22) Becker, O. M.; MacKerell, A. D. J.; Roux, B.; Watanabe, M. *Computational Biochemistry and Biophysics*; Marcel Dekker: New York, **2001**.

(23) Lii, J. H.; Allinger, N. L. *Journal of Computational Chemistry* **1991**, *12*, 186.

(24) Lii, J. H.; Allinger, N. L. *Journal of Computational Chemistry* **1998**, *19*, 1001.

(25) Maple, J. R.; Hwang, M. J.; Jalkanen, K. J.; Stockfisch, T. P.; Hagler, A. T. *Journal of Computational Chemistry* **1998**, *19*, 430.

(26) Sun, H. *Journal of Physical Chemistry B* **1998**, *102*, 7338.

(27) Halgren, T. A. *Journal of Computational Chemistry* **1996**, *17*, 490.

(28) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *Journal of the American Chemical Society* **1992**, *114*, 10024.

(29) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *Journal of Physical Chemistry* **1990**, *94*, 8897.

(30) Mackerell, A. D.; Wiorkiewiczkuczera, J.; Karplus, M. *Journal of the American Chemical Society* **1995**, *117*, 11946.

(31) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *Journal of Physical Chemistry B* **1998**, *102*, 3586.

(32) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *Journal of the American Chemical Society* **1984**, *106*, 765.

(33) Neria, E.; Fischer, S.; Karplus, M. *Journal of Chemical Physics* **1996**, *105*, 1902.

(34) Ceccarelli, M.; Marchi, M. *Journal of Physical Chemistry B* **1997**, *101*, 2105.

(35) Feig, M.; Pettitt, B. M. *Biophysical Journal* **1998**, *75*, 134.

(36) Venable, R. M.; Zhang, Y. H.; Hardy, B. J.; Pastor, R. W. *Science* **1993**, *262*, 223.

(37) Feller, S. E.; Yin, D. X.; Pastor, R. W.; MacKerell, A. D. *Biophysical Journal* **1997**, *73*, 2269.

(38) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *Journal of the American Chemical Society* **1996**, *118*, 11225.

(39) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. *Journal of Physical Chemistry* **1994**, *98*, 13077.

(40) Jorgensen, W. L.; Tiradorives, J. *Journal of the American Chemical Society* **1988**, *110*, 1657.

(41) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *Journal of the American Chemical Society* **1991**, *113*, 2810.

(42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins-Structure Function and Bioinformatics* **2006**, *65*, 712.

(43) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *Journal of Physical Chemistry B* **2001**, *105*, 6474.

(44) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; BIOMOS b.v.: Zurich, **1996**.

(45) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J Phys Chem B* **2007**, *111*, 7812.

(46) Roux, B.; Simonson, T. *Biophys Chem* **1999**, *78*, 1.

(47) Zhou, R. *Proteins* **2003**, *53*, 148.

(48) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24.

(49) Lee, M. S.; Salsbury, F. R., Jr.; Olson, M. A. *J Comput Chem* **2004**, *25*, 1967.

(50) Warshel, A.; Levitt, M. *J Mol Biol* **1976**, *103*, 227.

(51) van Gunsteren, W. F.; Karplus, M. *Macromolecules* **1982**, *15*, 1528.

(52) Ermak, D. L.; Mccammon, J. A. *Journal of Chemical Physics* **1978**, *69*, 1352.

(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926.

(54) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *In Intermolecular Forces*; Reidel: Dordrecht, **1981**.

(55) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *Journal of Physical Chemistry* **1987**, *91*, 6269.

(56) Ferrario, M.; Haughney, M.; Mcdonald, I. R.; Klein, M. L. *Journal of Chemical Physics* **1990**, *93*, 5156.

(57) Jorgensen, W. L.; Briggs, J. M.; Contreras, M. L. *Journal of Physical Chemistry* **1990**, *94*, 1683.

(58) Mennucci, B.; Martinez, J. M. *Journal of Physical Chemistry B* **2005**, *109*, 9818.

(59) Foloppe, N.; MacKerell, A. D. *Journal of Computational Chemistry* **2000**, *21*, 86.

(60) Langley, D. R. *Journal of Biomolecular Structure & Dynamics* **1998**, *16*, 487.

(61) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *Journal of Biomolecular Structure & Dynamics* **1999**, *16*, 845.

(62) Kirkwood, J. G.; Buff, F. P. *Journal of Chemical Physics* **1951**, *19*, 774.

(63) Bennaim, A. *Molecular theory of solutions*; Oxford University Press: New York, **2006**.

(64) Bennaim, A. *Journal of Chemical Physics* **1977**, *67*, 4884.

(65) Bentenitis, N.; Cox, N. R.; Smith, P. E. *Journal of Physical Chemistry B* **2009**, *113*, 12306.

(66) Chitra, R.; Smith, P. E. *Journal of Physical Chemistry B* **2002**, *106*, 1491.

(67) Gee, M. B.; Cox, N. R.; Jiao, Y. F.; Bentenitis, N.; Weerasinghe, S.; Smith, P. E. *Journal of Chemical Theory and Computation* **2011**, *7*, 1369.

(68) Gee, M. B.; Smith, P. E. *Abstracts of Papers of the American Chemical Society* **2008**, *236*.

(69) Gee, M. B.; Smith, P. E. *Journal of Chemical Physics* **2009**, *131*, 165101.

(70) Kang, M.; Smith, P. E. *J Comput Chem* **2006**, *27*, 1477.

(71) Kang, M.; Smith, P. E. *Fluid Phase Equilibria* **2007**, *256*, 14.

(72) Kang, M.; Smith, P. E. *Journal of Chemical Physics* **2008**, *128*.

(73) Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochemistry and Biophysics* **2008**, *50*, 1.

(74) Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilibria* **2010**, *290*, 43.

(75) Ploetz, E. A.; Smith, P. E. *Physical Chemistry Chemical Physics* **2011**, *13*, 18154.

(76) Ploetz, E. A.; Smith, P. E. *Journal of Chemical Physics* **2011**, *135*.

(77) Smith, P. E. *Journal of Physical Chemistry B* **2004**, *108*, 18716.

(78) Smith, P. E. *Journal of Physical Chemistry B* **2004**, *108*, 16271.

(79) Smith, P. E. *Journal of Physical Chemistry B* **2006**, *110*, 2862.

(80) Smith, P. E. *Biophysical Journal* **2006**, *91*, 849.

(81) Smith, P. E.; Mazo, R. A. *Journal of Physical Chemistry B* **2008**, *112*, 7875.

(82) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, *119*, 11342.

(83) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2003**, *107*, 3891.

(84) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, *118*, 5901.

(85) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, *118*, 10663.

(86) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2004**, *121*, 2180.

(87) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2005**, *109*, 15080.

(88) Marcus, Y. *Monatshefte Fur Chemie* **2001**, *132*, 1387.

(89) Ruckenstein, E.; Shulgin, I. *Journal of Physical Chemistry B* **1999**, *103*, 10266.

(90) Ruckenstein, E.; Shulgin, I. *Fluid Phase Equilibria* **2001**, *180*, 281.

(91) Shulgin, I.; Ruckenstein, E. *Journal of Physical Chemistry B* **1999**, *103*, 872.

(92) Shulgin, I.; Ruckenstein, E. *Industrial & Engineering Chemistry Research* **2002**, *41*, 6279.

(93) Shulgin, I.; Ruckenstein, E. *Polymer* **2003**, *44*, 901.

(94) Shulgin, I. L.; Ruckenstein, E. *Journal of Chemical Physics* **2005**, *123*, 054909.

(95) Shulgin, I. L.; Ruckenstein, E. *Journal of Physical Chemistry B* **2006**, *110*, 12707.

(96) Shulgin, I. L.; Ruckenstein, E. *Journal of Physical Chemistry B* **2007**, *111*, 3990.

(97) Shulgin, I. L.; Ruckenstein, E. *Fluid Phase Equilibria* **2007**, *260*, 126.

(98) Shulgin, I. L.; Ruckenstein, E. *Journal of Physical Chemistry B* **2008**, *112*, 3005.

(99) Shimizu, S. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 1195.

(100) Shimizu, S.; Boon, C. L. *Journal of Chemical Physics* **2004**, *121*, 9147.

(101) Shimizu, S.; Matubayasi, N. *Chemical Physics Letters* **2006**, *420*, 518.

(102) Shimizu, S.; McLaren, W. M.; Matubayasi, N. *Journal of Chemical Physics* **2006**, *124*, 234905.

(103) Shimizu, S.; Smith, D. J. *Journal of Chemical Physics* **2004**, *121*, 1148.

(104) Hall, D. G. *Journal of the Chemical Society-Faraday Transactions* **1991**, *87*, 3523.

(105) Zielenkiewicz, W.; Kulikov, O. V.; Krestov, G. A. *Bulletin of the Polish Academy of Sciences-Chemistry* **1992**, *40*, 293.

(106) Matteoli, E.; Lepori, L. *Journal of Chemical Physics* **1984**, *80*, 2856.

(107) Matteoli, E.; Lepori, L. *Journal of the Chemical Society-Faraday Transactions* **1995**, *91*, 431.

(108) Ellegaard, M. D.; Abildskov, J.; O'Connell, J. P. *Fluid Phase Equilibria* **2011**, *302*, 93.

(109) Wedberg, R.; O'Connell, J. P.; Peters, G. H.; Abildskov, J. *Molecular Simulation* **2010**, *36*, 1243.

(110) Kang, M.; Smith, P. E. *Journal of Computational Chemistry* **2006**, *27*, 1477.

(111) Guha, A.; Mukherjee, D. *Journal of the Indian Chemical Society* **1997**, *74*, 195.

(112) Guha, A.; Ghosh, N. K. *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **1998**, *37*, 97.

(113) Imai, T.; Kinoshita, M.; Hirata, F. *Journal of Chemical Physics* **2000**, *112*, 9469.

(114) Imai, T.; Harano, Y.; Kovalenko, A.; Hirata, F. *Biopolymers* **2001**, *59*, 512.

(115) Imai, T.; Takahiro, T.; Kovalenko, A.; Hirata, F.; Kato, M.; Taniguchi, Y. *Biopolymers* **2005**, *79*, 97.

(116) Lynch, G. C.; Perkyns, J. S.; Pettitt, B. M. *Journal of Computational Physics* **1999**, *151*, 135.

(117) Matteoli, E.; Mansoori, G. A. *Journal of Chemical Physics* **1995**, *103*, 4672.

(118) Matteoli, E. *Journal of Molecular Liquids* **1999**, *79*, 101.

(119) Nain, A. K. *Journal of Solution Chemistry* **2008**, *37*, 1541.

(120) Pandey, J. D.; Verma, R. *Chemical Physics* **2001**, *270*, 429.

(121) Patil, K. J.; Mehta, G. R.; Dhondge, S. S. *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **1994**, *33*, 1069.

(122) Pjura, P. E.; Paulaitis, M. E.; Lenhoff, A. M. *Aiche Journal* **1995**, *41*, 1005.

(123) Rosgen, J.; Pettitt, B. M.; Bolen, D. W. *Protein Science* **2007**, *16*, 733.

(124) Rosgen, J. *Osmosensing and Osmosignaling* **2007**, *428*, 459.

(125) Schellman, J. A. *Quarterly Reviews of Biophysics* **2005**, *38*, 351.

(126) Warshavsky, V. B.; Song, X. Y. *Physical Review E* **2008**, *77*.

(127) Allen, M. P. T., D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, **1989**.

(128) Widom, B. *Statistical Mechanics: A Concise Introduction for Chemists*; Cambridge University Press: New York, **2002**.

(129) McQuarrie, D. A. *Statistical Mechanics*; Dover Publications: New York, **1976**.

(130) Cramer, C. *Essentials of Computational Chemistry: Theories and Models, Second ed.*; John Wiley & Sons Ltd.: West Sussex, **2004**.

(131) Brooks, C. L. *Journal of Molecular Biology* **1992**, *227*, 375.

(132) Tiradorives, J.; Jorgensen, W. L. *Biochemistry* **1993**, *32*, 4175.

(133) Roccatano, D.; Amadei, A.; Di Nola, A.; Berendsen, H. J. C. *Protein Science* **1999**, *8*, 2130.

(134) Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1993**, *232*, 600.

(135) Mark, A. E.; Vangunsteren, W. F. *Biochemistry* **1992**, *31*, 7745.

(136) Pande, V. S.; Rokhsar, D. S. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 9062.

(137) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(138) Tsai, J.; Levitt, M.; Baker, D. *Journal of Molecular Biology* **1999**, *291*, 215.

(139) Perera, A.; Sokolic, F. *Journal of Chemical Physics* **2004**, *121*, 11272.

(140) Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C. *J Comput Chem* **2003**, *24*, 21.

(141) Mu, Y. G.; Kosov, D. S.; Stock, G. *Journal of Physical Chemistry B* **2003**, *107*, 5064.

(142) Mazur, A. K. *Journal of the American Chemical Society* **2003**, *125*, 7849.

(143) Gao, J. L. *Journal of Physical Chemistry B* **1997**, *101*, 657.

(144) Xie, W.; Gao, J. *J Chem Theory Comput* **2007**, *3*, 1890.

(145) Xie, W.; Pu, J.; Mackerell, A. D.; Gao, J. *J Chem Theory Comput* **2007**, *3*, 1878.

(146) Yu, H. B.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D.; Roux, B. *Journal of Chemical Theory and Computation* **2010**, *6*, 774.

(147) Piquemal, J. P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J Chem Phys* **2006**, *124*, 104101.

(148) Cisneros, G. A.; Piquemal, J. P.; Darden, T. A. *Journal of Chemical Physics* **2006**, *125*.

(149) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Journal of Computational Chemistry* **1998**, *19*, 535.

(150) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983**, *4*, 187.

(151) Best, R. B.; Buchete, N. V.; Hummer, G. *Biophysical Journal* **2008**, *95*, 4494.

(152) Cacace, M. G.; Landau, E. M.; Ramsden, J. J. *Quarterly Reviews of Biophysics* **1997**, *30*, 241.

Chapter 2 - A Kirkwood-Buff Derived Force Field for Alcohols in Water

2.1 Abstract

Computer simulation is used to study the properties of a series of alcohols and water mixtures in an attempt to develop and validate a force field for alcohols specifically designed to accurately reproduce the experimentally derived Kirkwood-Buff integrals (KBIs). Mixtures covering the entire composition range where investigated for methanol, ethanol, n-propanol, n-butanol, i-propanol, t-butanol, and n-octanol. The results suggest that, to a very good approximation, the parameters developed for the simulation of methanol and water mixtures also provide a reasonable description of other primary alcohol and water mixtures. The results indicate that the hydroxyl and hydrocarbon group parameters appear to be both transferrable and additive among primary alcohols. However, in the case of the secondary alcohol solutions, alternative parameters were required for oxygen. Other properties of the solution mixtures besides the KB integrals were also well reproduced including the enthalpy of mixing, translational diffusion constants, and dielectric properties.

2.2 Introduction

Mixtures of alcohols and water represent model systems of interest to a variety of researchers. Generally, systematic changes in characteristics such as structural, hydrogen binding, thermodynamic, dielectric properties, etc are observed with increasing chain length.[1-3] For instance, alcohol and water mixtures can be considered as one of the simplest systems in which to study the hydrophobic effect. As the length of chain increases, the solubility of alcohols in water decreases, due to aggregation of aliphatic groups.[4-7] Methanol, ethanol and n-propanol are soluble in aqueous solution over the full composition range. However, n-butanol displays a phase separation in water with miscibility gap between 0.02 and 0.49 mol fraction at 293K. Furthermore, the behavior of longer chain alcohols such as n-octanol is also of interest even though it is immiscible with water. A fundamental use of n-octanol is found in the pharmacotherapy and drug delivery fields. The logarithm of the n-octanol/water partition coefficient, which is also expressed as lipophilicity (logP), is the most important indicator for blood-brain barrier (BBB) penetration, and whether drugs can be delivered to the central nervous

system through the BBB. In computer simulations, n-octanol has also been widely used as a model for membrane bilayers, with the polar head groups outside facing the solvent and the nonpolar carbon chains in the middle away from solvent.

Aqueous mixtures of simple alcohols, such as methanol and ethanol, have been studied in detail both experimentally and theoretically.[1,8-13] Experimental studies include X-ray scattering, diffusion constant, dielectric properties, etc. In terms of simulation, there are many studies using both Monte Carlo (MC) and Molecular Dynamics (MD) techniques. Studies of water-alcohol mixtures by MC simulations include a series of molecular properties such as radial distribution function, and bulk properties such as solubility, density, surface tension and interactions at infinite dilution as well,[14-17] all of which are based on stable systems. In contrast, MD simulation provided more information about dynamics such as velocity and properties derived from it.

There are several different force fields developed for the simulation of alcohols. These include the CHARMM,[18] Amber,[19] OPLS,[20] and GROMOS[21] parameter sets with different sets of parameters for the alcohol models. Using t-butanol as an example, Gromos[21] and OPLS[20] published the parameter of the alcohol (based on methanol) in 2000 and 1986, respectively. The Siepmann group[22] have developed a TraPPE force field for phase equilibria studies based on Monte Carlo simulation in 2008. Different force fields developed using different approaches will often display a range of advantages and disadvantages when applied to real systems. In particular, the accuracy of the simulation results is strongly linked to the accuracy of the force field describing the intermolecular and intramolecular interactions.[10,20,23] In an effort to provide more accurate models for computer simulation, Weerasinghe and Smith[24] developed a force field for methanol using the KB theory of solutions as a guide. Lee and Vegt[23] have taken a similar approach to improve the parameters for t-butanol and water mixtures. However, a general force field developed (or tested) using KB theory has not been provided.

Most force fields mentioned above are non-polarizable force fields. To more accurately express the molecular behavior several polarizable force fields have also been developed. For example, polarizable force fields based on electronic structural theory, induced dipoles, point charges, distributed multipoles, bond polarization theory, etc, have been suggested.[25-30] Common polarizable force fields for alcohols include the CHARMM polarizable force field developed by Patel[31], the CHARMM polarizable force fields based on the classical Drude oscillator developed by MacKerell,[32] and the AMBER polarizable force field developed by Caldwell and Kollman.[33]

In order to validate the development of force fields, people studies potential energy function of alcohols in aqueous solution;[5] the effect of hydrophobic association before and after adding salt into alcohol solution.[34]

Although many force fields for alcohols exist in the literature, we have observed numerous issues with such models that have not been thoroughly tested against properties of aqueous mixtures.[6,25,27] Therefore, improved force fields providing better agreement with the available experimental data for a large range of aqueous mixtures are still highly desired. Recently, we have been developing a force field (KBFF) for molecular dynamics (MD) simulations which is specifically designed to reproduce Kirkwood-Buff (KB) integrals obtained from the experimental data on solutions mixtures.[24,35,36] KB integrals have been shown to be a sensitive probe of the molecular distributions observed for different solutions.[24,35-46] The KBFF method is primarily aimed at providing accurate force fields for the simulation of peptides and proteins. As several amino acids contain the hydroxyl group it is necessary to develop models for the alcohols in a manner consistent with our previous systems. In this study, we extend our previously developed KBFF methanol model[24] to include a series of primary alcohols - such as ethanol, n-propanol, n-butanol and n-octanol – together with several branched alcohols such as i-propanol and t-butanol. The proposed force field is designed for use with the SPC/E water model.[47] Computer simulations are performed to examine the properties of alcohol and water mixtures in an attempt to validate the force field models. In particular, we study the thermodynamic properties and aggregation behavior at atomic level. Mixtures covering the entire composition range are examined. These alcohol models can be modified as a basis of peptide side chain for our future work.

2.3 Theory

Kirkwood-Buff theory is a well established theory of solutions which has been utilized to provide information concerning the local distributions of the various species in a solution mixture. The current approach uses KB theory as tool to provide target data for the development and testing of force field models. KB theory is described in detail elsewhere.[48,49] Here, we briefly outline the basic theory and approach. KB theory provides a link between the thermodynamic properties of a solution mixture and integrals over the radial distribution functions between the

various solution components. The KB integrals (KBI) between the different solution components are defined by,

$$G_{ij} = 4\pi \int_0^\infty [g_{ij}^{\mu VT}(r) - 1] r^2 dr \tag{2.1}$$

where G_{ij} is the KBI between species i and j, $g_{ij}^{\mu VT}$ is the corresponding center of mass based radial distribution function (rdf) in the Grand Canonical (μVT) ensemble, and r is the center of mass to center of mass distance. Combinations of KBIs and number densities $\rho_i = N_i/V$ then provide expressions for thermodynamic properties of the solution in any ensemble. Of particular interest are the isothermal compressibility (κ_T), the partial molar volumes ($\overline{V}_i$), and derivatives of the chemical potentials (μ_i) as shown below.

One of the advantages of using KB theory is that it provides a more physical picture of the relative molecular distributions in solution at the composition of interest. To do this one can define a series of excess coordination numbers, $N_{ij} = \rho_j G_{ij}$, which quantify the excess number of j particles around a central i particle in a given local volume of solution over the number of j particles found in the same volume of bulk solution. A value of N_{ij} much greater than zero indicates an excess of species j in the vicinity of species i (over a random distribution), presumably due to some net favorable interactions, while a large negative value corresponds to a depletion of species j surrounding the i species. KB theory can also be expressed in terms of particle-particle fluctuations within a local (open) region of a solution mixture. The relationship between the KBIs or excess coordination numbers and the corresponding fluctuating properties is given by

$$B_{ij} = \rho_i(\delta_{ij} + N_{ij}) = \frac{< \delta N_i \delta N_j >}{V} \tag{2.2}$$

where $\delta N_i = N_i - \langle N_i \rangle$, δ_{ij} is the Kronecker Delta function, and V is the volume of interest. Hence, it is also referred to as the Fluctuation Theory (FT) of solutions.

More recently, the approach has been extended to include particle-energy and energy-energy fluctuations characterizing local regions of a solution.[50,51] These additional fluctuating properties allow one to access additional thermodynamic properties such as excess partial molar enthalpies ($\overline{H}_i^E$), the isobaric thermal expansion coefficient (α_P), and the constant pressure molar

heat capacity ($C_{P,m}$), all expressed in terms of local fluctuations. To do this one requires two additional fluctuating properties

$$\rho_i F_i = \frac{< \delta N_i \delta \varepsilon >}{V} \tag{2.3}$$

and

$$\rho \Delta_m = \frac{< \delta \varepsilon \delta \varepsilon >}{V} \tag{2.4}$$

where ρ is the total number density, $\varepsilon = E - \sum_i N_i H_i^o$, E is the instantaneous local internal energy, H_i^o is the enthalpy per molecule of species *i* in the pure liquid, and the sum is performed over all the solution components. We note that the above definitions are slightly different to those presented previously.[50] The expressions provided here are equivalent to the previous set, but are somewhat simpler due to the slightly different definition of ε where we have used the pure enthalpies instead of energies.

The above fluctuating quantities can be obtained from computer simulations. Particle-particle fluctuations can be determined directly via Equation 2.1 or 2.3 for any type of interaction potential. However, particle-energy and energy-energy fluctuations require the use of pairwise additive potentials to be of practical use. KB theory provides exact expressions for a series of thermodynamic properties in terms of the above fluctuating quantities. The relevant expressions for a binary mixture of a solvent (1) and solute (2) at a mole fraction composition (x_i) are

$$\mu_{ij} = \beta \left(\frac{\partial \mu_i}{\partial x_j} \right)_{T,P} = (2\delta_{ij} - 1) \frac{\rho}{x_i \eta} \tag{2.5}$$

$$\overline{V}_i = \frac{1 + N_{jj} - N_{ij}}{\eta} \tag{2.6}$$

$$\overline{H}_i^E = \rho \overline{V}_i H_m^E + \frac{\rho_j}{\eta} \left(F_i - F_j \right) \tag{2.7}$$

$$\boldsymbol{RT\kappa_T = \frac{\xi}{\eta}}$$

(2.8)

$$RT^2\alpha_P = \rho RT\kappa_T H_m^E - \phi_1 F_1 - \phi_2 F_2$$

(2.9)

$$RT^2 C_{P,m} = \Delta_m + RT^2\alpha_P H_m^E - x_1 F_1 \overline{H}_1^E - x_2 F_2 \overline{H}_2^E$$

(2.10)

where β = 1/RT, R is the Gas constant, $\phi_i = \rho_i \bar{V}_i$ is the volume fraction of *i*, H_m^{E} is the excess enthalpy of mixing, and $\xi = (1 + N_{11})(1 + N_{22}) - N_{12}N_{21}$ with $\eta = \rho_2(1 + N_{11} - N_{21}) + \rho_1(1 + N_{22} - N_{12})$. In principle, every quantity appearing on the right hand side of the above equations can be determined from a single simulation at a specific composition, given the properties of the pure components. The complexity of the above expressions illustrates the usefulness of the local fluctuations for the analysis of simulation results. For instance, if one were to determine a simulated partial molar volume or excess enthalpy that disagreed with experiment, it would be extremely difficult to relate this error to the force field parameters. However, this is much more likely to be possible, though still not trivial, if one observes an error in N_{ij} indicating an increase or decrease in the association of *i* and *j* particles. It is this aspect of the combined use of KB theory and computer simulation, together with the computationally efficient access to chemical potential (activity) derivatives, which represent the major strengths of the current approach.

Alternatively, if one knows all the relevant experimental data for a binary mixture as a function of solution composition, one can invert the whole procedure and express the fluctuating quantities in terms of the experimental data to provide,

$$\delta_{ij} + N_{ij} = \rho_j RT\kappa_T + \frac{1}{x_1 x_2}\frac{\rho_j}{\rho_i}\frac{(1-\phi_i)(1-\phi_j)}{\mu_{ij}}$$

(2.11)

$$F_i = -RT^2\alpha_P + (\delta_{i1} + N_{i1})\overline{H}_1^E + (\delta_{i2} + N_{i2})\overline{H}_2^E$$

(2.12)

$$\Delta_m = RT^2 C_{P,m} - RT^2\alpha_P H_m^{\mathrm{E}} + x_1 F_1 \overline{H}_1^E + x_2 F_2 \overline{H}_2^E$$

(2.13)

which can be solved in a stepwise manner. As the above expressions are exact this type of analysis provides experimental data in a form which is easy to access via simulation and can be

used as target data to help determine and improve the quality of a force field. This is the approach we have taken here. While many of the experimental properties used to extract the fluctuating quantities can also be obtained directly from a simulation by other means - the partial molar volumes and excess enthalpies for example - other properties such as derivatives of the chemical potentials are much more difficult to obtain from conventional methods. Furthermore, the above fluctuating properties provide a much simpler physical picture of local solution behavior compared to the corresponding thermodynamic properties.

2.4 Method

2.4.1 KBFF Models and Parameter Development

The KBFF models used and developed in this study are simple classical nonpolarizable united atom models. The future aim of the force field is to provide a computationally efficient scheme for biomolecular simulations which is compatible with existing force fields and codes developed for the simulation of biomolecules. The nonbonded interactions are treated by the usual Lennard-Jones (LJ) 6-12 plus Coulomb potential, which is the most commonly used potential for biomolecular simulation. This is augmented with the typical bonded interactions. The total potential energy is given by the following terms

$$V^{nonbonded} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.14}$$

$$V^{bond} = \sum \tfrac{1}{2} k_r (r - r_o)^2 \tag{2.15}$$

$$V^{angle} = \sum \tfrac{1}{2} k_\theta (\theta - \theta_o)^2 \tag{2.16}$$

$$V^{torsion} = \sum k_\phi [1 + \cos(n\phi - \phi_o)] \tag{2.17}$$

$$V^{improper} = \sum \tfrac{1}{2} k_\xi (\xi - \xi_o)^2 \tag{2.18}$$

where all the symbols have their usual meaning. The LJ parameters for oxygen and hydrogen were taken from our previous study of methanol and water mixtures.[24] The united atom carbon LJ parameters were taken directly from the literature.[52] All LJ and Coulomb interactions between 1-4 atoms were excluded, and geometric combination rules were used for both the σ and ε parameters. The equilibrium molecular geometry (bond and angle terms) was taken from the OPLS force field,[20] with force constants from the GROMOS force field. The improper dihedral terms were taken from the GROMOS force field.[52] The force constants for the torsional potentials were obtained by fitting the rotational energies of small hydrocarbons and alcohols obtained from literature quantum mechanical calculations.[20] The SPC/E water model was used for all simulations.[47] The primary parameters of interest in the KBFF models are the partial atomic charges. Typically, the KBFF models are parameterized by variation of the charge distribution to best reproduce the experimentally derived KBIs as a function of solution composition.[40-42,44] This is the approach used to study aqueous methanol solutions.[24] It was found that the methanol charges developed previously also worked well for the other linear chain alcohols studied here. Consequently, new charge distributions were tested until reasonable agreement with experiment was obtained. Finally, two alcohol mixtures which were not used in the parameterization studies were simulated as an independent test of the final models.

Table 2.1 KBFF non-bonded force field parameters.

Model	atom	ε, kJ/mol	σ, nm	q, |e|
Alcohols				
KBFF	O	0.6506	0.3192	-0.82/-0.902
	H	0.0880	0.1580	0.52/0.572
	CH_3	0.8672	0.3748	0.30
	CH_2	0.4105	0.4070	0
	CH_1	0.0949	0.5019	0.33
	C	0.4170	0.3770	0.30
Water				
SPC/E	O	0.6506	0.3166	-0.8476
	H	0	0	0.4238

The final non-bonded force field parameters are presented in Table 2.1. The bonded parameters are provided in the Supporting Information Table 2.3.SPC/E parameters were taken from ref 2. Geometric combination rules were used for both the σ and ε parameters. Non-bonded 1-4 interactions were excluded.

2.4.2 Fluctuation Analysis of the Experimental Data

In order to compare the simulated and experimentally derived fluctuations one has to extract the target data from existing experimental properties. To obtain the KBIs one requires composition dependent chemical potentials (activities), partial molar volumes, and isothermal compressibilities. To obtain the F's one also requires the excess enthalpy of mixing and thermal expansion data, while the determination of $\mathbf{f}_m$ then requires the heat capacity. Excess molar Gibbs free energies, enthalpies and volumes at 298.15 K and 1 bar for alcohol and water mixtures were taken from the literature.[1,3,9,53-55] To obtain the corresponding partial molar quantities the experimental data was first fitted using the standard Redlich-Kister equation,[56]

$$X_m^E = x_1 x_2 \sum_i a_i (x_1 - x_2)^i \tag{2.19}$$

where a_i are fitting constants and X is the molar Gibbs free energy (G), enthalpy (H), or volume (V). The number of terms in the summation was typically i = 0 through 6. After fitting the raw experimental data the corresponding excess partial molar quantities at any composition can be obtained from the thermodynamic relationship,

$$\overline{X}_i^E = X_m^E + (1 - x_i)\left(\frac{\partial X_m^E}{\partial x_i}\right)_{P,T} \tag{2.20}$$

which provides the required excess partial molar volumes, enthalpies, and chemical potentials. The chemical potential derivatives can be obtained from the second derivative of the excess Gibbs free energy,

$$x_2 \mu_{22} = 1 + x_1 x_2 \beta \left(\frac{\partial^2 G_m^E}{\partial x_2^2}\right)_{T,P} \tag{2.21}$$

and the Gibbs-Duhem expression $x_1\ d\mu_1 + x_2\ d\mu_2 = 0$. The required compressibilities, thermal expansions, and heat capacities were taken directly from the literature and fitted to Equation 2.19. The resulting data are in agreement with previous determinations of the excess and partial molar volumes of alcohol and water,[55] and previous determinations of the KB integrals for regions where the KBIs are statistically reliable.[24,50]

2.4.3 Molecular Dynamics Simulation and General Analysis

Molecular dynamics simulations were performed using the KBFF models together with the SPC/E water model as implemented in the GROMACS 3.3.3 package.[47] All simulations were performed in the isothermal isobaric ensemble at 300 K and 1 atm using the weak coupling technique[57] to modulate the temperature and pressure with relaxation times of 0.1 and 0.5 ps, respectively. A time-step of 2 fs was used and the bond lengths were constrained using the Lincs (alcohols) and Settle (water) algorithmns. The particle mesh Ewald technique was used to evaluate electrostatic interactions.[58] A real space convergence parameter of 3.5 nm^{-1} was used in combination with twin range cutoffs of 0.8 and 1.5 nm, and a nonbonded update frequency of 10 steps. Random initial configurations of molecules in a cubic box were used. Initial configurations of the different solutions were generated from a cubic box ($L \approx 6.0$ nm) of equilibrated water molecules by randomly replacing waters with alcohol until the required concentration was attained. The steepest descent method was then used to perform 100 steps of minimization. This was followed by extensive equilibration, which was continued until the rdfs displayed no drift with time (typically 15 ns). Total simulation times were in the 20-35 ns range, and the final 15-30 ns were used for calculating ensemble averages. Configurations were saved every 0.1 ps for the calculation of various properties. A summary of all the systems and compositions studied here are provided in the Supporting Information Table 2.4. Translational self-diffusion constants (D_i) were determined using the mean square fluctuation approach,[10] relative permittivities (ϵ) from the dipole moment fluctuations,[59] and excess enthalpies of mixing (H_m^E) from the average potential energies.[21] Errors ($\pm 1\sigma$) in the simulation data were estimated by using five or six block averages.

2.4.4 Fluctuation Analysis of the Simulated Data

The KB integrals obtained from an analysis of the experimental data correspond to integrals over rdfs in the μVT ensemble. The infinite limit cannot be reached in practice because

one typically has a finite system which is usually closed with respect to each particle. However, one can reasonably approximate the required KBIs by assuming that,

$$G_{ij} = 4\pi \int_0^{\infty} [g_{ij}^{\mu VT}(r) - 1] r^2 dr \approx 4\pi \int_0^{R} [g_{ij}^{NPT}(r) - 1] r^2 dr \tag{2.22}$$

where *R* is a cutoff distance beyond which the rdfs are essentially unity, *i.e.* the bulk solution values. In practice, this condition is difficult to achieve precisely unless one uses very large systems. However, a reasonable approximation is to determine $G_{ij}(R)$ and average the values over a short distance range, typically one molecular diameter.[36,48] In this work the final KB integral values have been obtained by averaging between 1.50 and 2.00 nm.

A slightly different approach was adopted for the particle-energy (F_i) and energy-energy (Δ_m) fluctuating quantities. To evaluate these properties one has to determine the energy of a given volume of the solution. This is possible if one can assign energies to each particle in the system. Single particle potential energies were evaluated by analyzing the trajectory and recalculating the intermolecular potential energy of the system using a simple group based Coulomb plus LJ expression with a 1.5 nm cutoff distance. One half of the potential energy for the interaction between each pair of molecules was then assigned to each molecule for each configuration. All intramolecular potential energies were assumed to be the same as observed for the pure liquids and were therefore neglected. A series of random origins were then chosen and the instantaneous energy and number of particles determined as a function of the volume of a local region centered on each origin. The fluctuations observed after averaging over the trajectory provide values of F_i and Δ_m as a function of local volume (sphere radius). A graph of F_i or Δ_m as a function of sphere radius should tend to a constant value for radii large enough to approach bulk solution distributions. The kinetic energy contribution to the local energy cancels in the calculation of the F's and one can use just the potential energies for both the instantaneous internal energies (E) and the reference pure solvent values. The kinetic energy contribution to the $\beta^2\Delta_m$ values does not cancel. For the nonlinear molecules studied here this contribution is simply 3 and was added to the simulated values. No quantum mechanical or vibrational contributions to the F_i or Δ_m values were determined for the rigid models used here. For comparison, we also determined the particle-particle fluctuations using the fluctuation type of approach. This has the advantage over the standard integration approach (Equation 2.22) in that one can choose

essentially an unlimited set of random origins for averaging, rather than be restricted to the set of molecule origins, but at the expense of a clearer physical interpretation.

2.5 Results and Discussion

KB theory relates integrals over distribution functions to the solution thermodynamics. When combined with computer simulation, however, this can be extended further by providing insight into contributions to the thermodynamics from the various solvation shells surrounding each species. This is observed in Figure 2.1 where we display the rdfs between solutes for the solutions studied here (other rdfs are provided in the supporting information) using the final optimized force field parameters provided in Table 2.1. The center of mass based rdfs provide a less detailed description of the solution mixture compared to the atom based rdfs, but they are most the most relevant distributions as far as the solution thermodynamics is concerned as described by KB theory. As indicated in Equation 2.22 all peaks above unity will contribute positively to the corresponding G_{ij} values, while all troughs below unity in the rdfs will contribute negatively to the final integral. Figure 2.1 clearly indicates that one would expect contributions over many solvation shells to distances of 1.5 nm or more. Sometimes the longer range oscillations cancel, an example being water and methanol, while for most other systems small changes in the rdfs at larger distances provided a significant contribution to the final G_{ij} values.

There are certain trends and features in the rdfs that deserve some discussion. Obviously, the peak positions are shifted to larger distances and broadened as the chain length for the n-alcohol increases. The changes with composition are relatively small for all the mixtures. This is even true for the alcohol rdf in n-BuOH at a composition (x_2 = 0.25) that phase separates. A visual inspection of this composition clearly indicates the separation into two phases, but this is less obvious from the alcohol rdf. It is only the long distance behavior, where the rdf stays consistently above unity, that captures this effect. The excess coordination numbers in this composition region are very large (N_{22} = 15, N_{11} = 150 and N_{21} = -80), but are not converged due to the immiscibility. All other rdfs were essentially unity beyond 1.5 nm. The secondary and tertiary alcohol mixtures general displayed more structure in their rdfs. The methanol-methanol rdf in both water and i-PrOH displayed the same features, but the first peak is much larger in the i-PrOH mixture. This is somewhat surprising as the enthalpies of mixing, for instance, indicate

that the i-PrOH/MeOH mixtures are much more ideal than the MeOH/HOH mixtures.[60] In comparison, the alcohol oxygen-oxygen rdfs display much less variation in peak position as indicated in Figure 2.2. However, significant changes to the peak height are observed for the first solvation shell. The first peak height increases as the hydrocarbon chain is extended or branched. Again, the distribution obtained for the i-PrOH/MeOH mixtures appears much more structured than that observed for MeOH/HOH mixtures. Additional oxygen-oxygen (solute and solvent) rdfs are provided in the Supplementary Material.

Figure 2.1 Center of mass based radial distribution functions (g_{22}) as a function of distance (nm) and alcohol mole fraction. Only selected mole fractions of 0.25 (black), 0.50 (red) and 0.75 (green) are displayed for clarity.

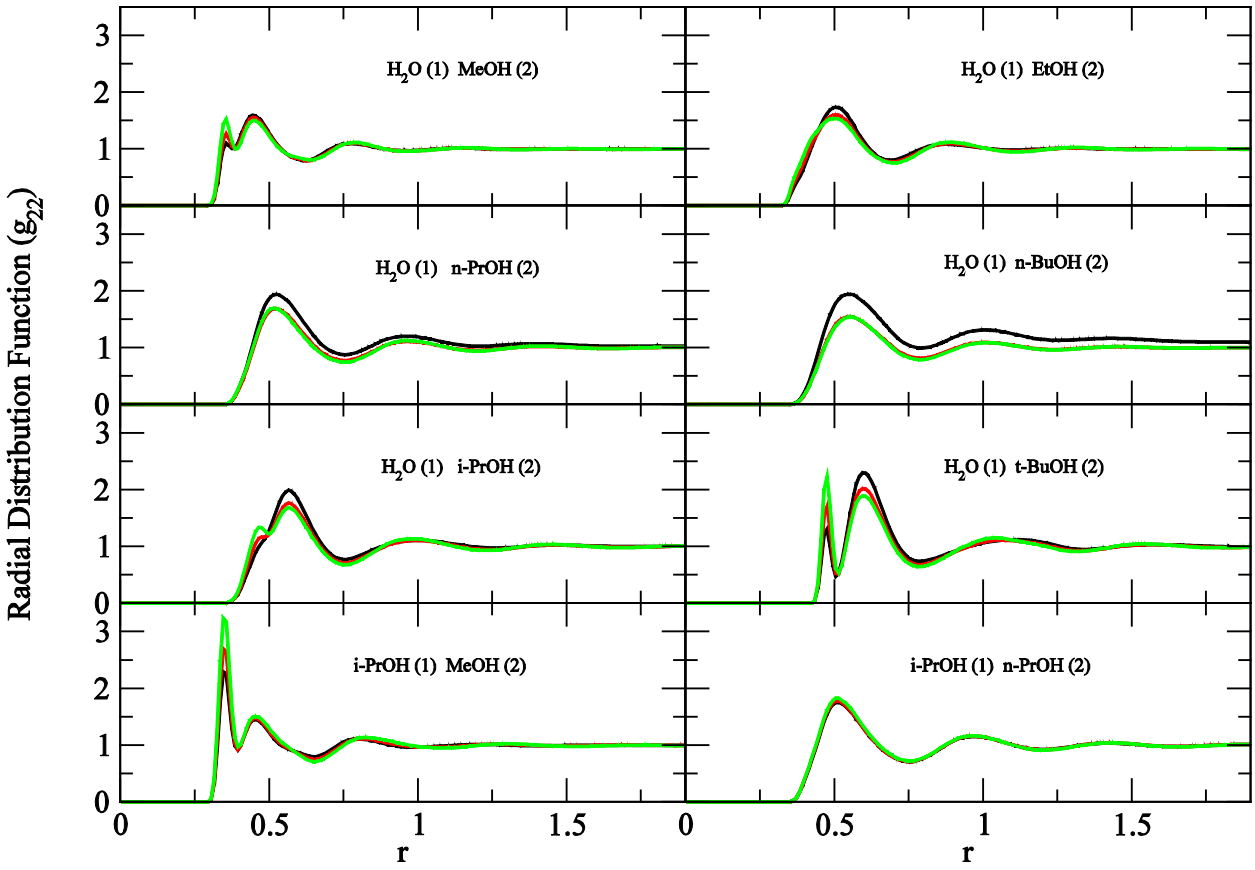
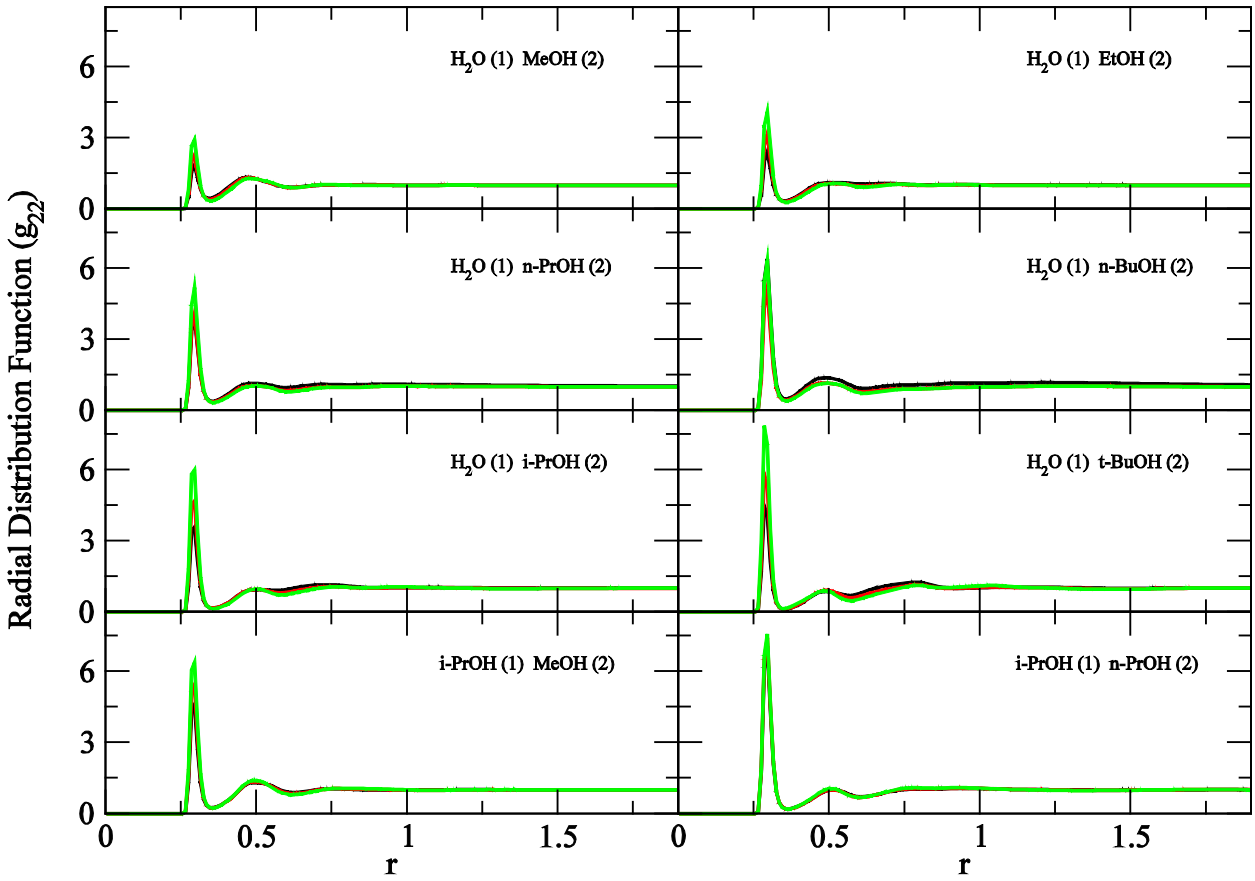


Figure 2.2 Oxygen-oxygen radial distribution functions (g_{22}) as a function of distance (nm) and alcohol mole fraction. Only selected mole fractions of 0.25 (black), 0.50 (red) and 0.75 (green) are displayed for clarity.



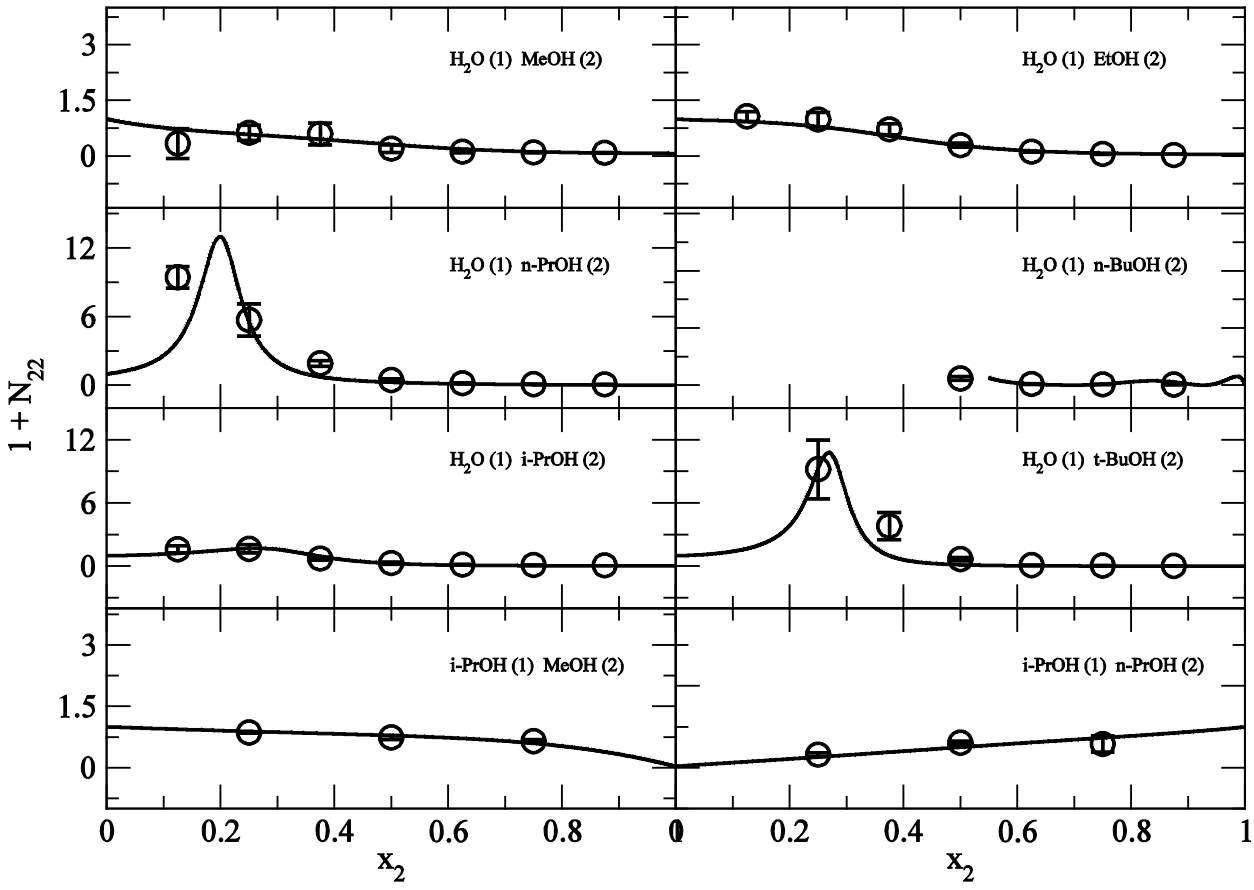
The experimental and simulated KB integrals are compared in Figure 2.3 as excess coordination numbers ($N_{ij} = \rho G_{ij}$). The use of excess coordination numbers helps to suppress the inherent uncertainties in both the experimental and simulated G_{ij} integrals at low j concentrations. Other excess coordination numbers are provided in the Supplementary Material and show similar agreement with experiment. The experimental data indicates larger values of N_{22} in the regions of x_2 = 0.20-0.25 as the alkyl chain increases. This results in a phase

separation for n-BuOH, and large positive values of N_{22} for n-PrOH and also t-BuOH. These large values describe significant self-aggregation between alcohol molecules at these compositions and suggest that the local regions in these solutions would consist fluctuate between regions containing a large number of alcohol or water molecules compared to the bulk distributions. Obtaining agreement between the simulated and experimental values of the KBIs, and therefore the local solution behavior, is the major aim of the KBFF models. The trends in the experimental data were well reproduced by the simulations. In particular, the large increases in $N_{22,}$ indicating an increase in self association, in the region of x_2 = 0.20-0.25 appear to be well reproduced. There was essentially quantitative agreement for N_{22} (N_{21} and N_{11}) over the composition range from 0.125 to 0.875. The only exceptions were the x_2 = 0.125 compositions for i-PrOH and t-BuOH. However, the KB analysis suggested that at this composition the KBIs for these solutions were not converged. This convergence could be improved by the use of larger systems sizes. The current systems sizes (L = 6 nm) are not small and hence this suggests rather large systems are required to fully capture the behavior of these two systems at low alcohol mole fractions.

Our initial investigations used the same atomic charge distribution developed for methanol for all the alcohols studied here. However, while this worked well for the linear alcohols, the results for the branched alcohols were relatively poor. This is to be expected as the effective charge distributions are affected by the local environment through polarization effects. The hydroxyl groups in linear alcohols can be considered, to a good approximation, to experience the same polarization effects due to the presence of hydrogen bonded water molecules and other alcohols. However, for branched alcohols this argument does not hold. The presence of additional methyl groups around the central hydroxyl group will tend to bias the polarization effects to one direction. This in turn will tend to a larger polarization. Hence, we tried to improve the secondary and tertiary alcohols by simply scaling the methanol charges (the simplest approach possible) to mimic the expected polarization changes. A 10% increase was observed to produce the best overall results. Interestingly, a larger (15-20%) scaling improved agreement for the KBIs, but had the undesirable side effect that the pure liquids appeared to behave as amorphous solids with very low (negligible) diffusion constants and relative permittivities. Hence, a smaller scaling factor was adopted. The agreement between the simulated and experimental KBIs was very reasonable using this approach. It should be noted

that most force fields do not reproduce the KBIs accurately, and that the errors can be substantial.[24,37]

Figure 2.3 Solute-solute excess coordination numbers ($N_{22} = \rho_2 G_{22}$) as a function of alcohol mole fraction. The solid black lines correspond to an analysis of the experimental data,[2,53,55,61,62] while the symbols correspond to the simulation data.



Our primary goal was to reproduce the experimental KBIs for mixtures of alcohols with water. This could be achieved by simply transferring the methanol charge distribution to other linear alcohols. Branched alcohols required modified charge distributions. As a further test of the transferability of both charge distributions we performed simulations of two additional alcohol

mixtures, i-PrOH with MeOH and i-PrOH with n-PrOH. A comparison to the experimental KBIs is very good and suggests that the models can be used for general alcohol mixtures.

Figure 2.4 The excess molar enthalpy of mixing (ΔH_m^E) in kJ/mol as a function of alcohol mole fraction. The solid lines correspond to the experimental data,[1,3,9,12,54,63-66] and the circles to the simulation data.

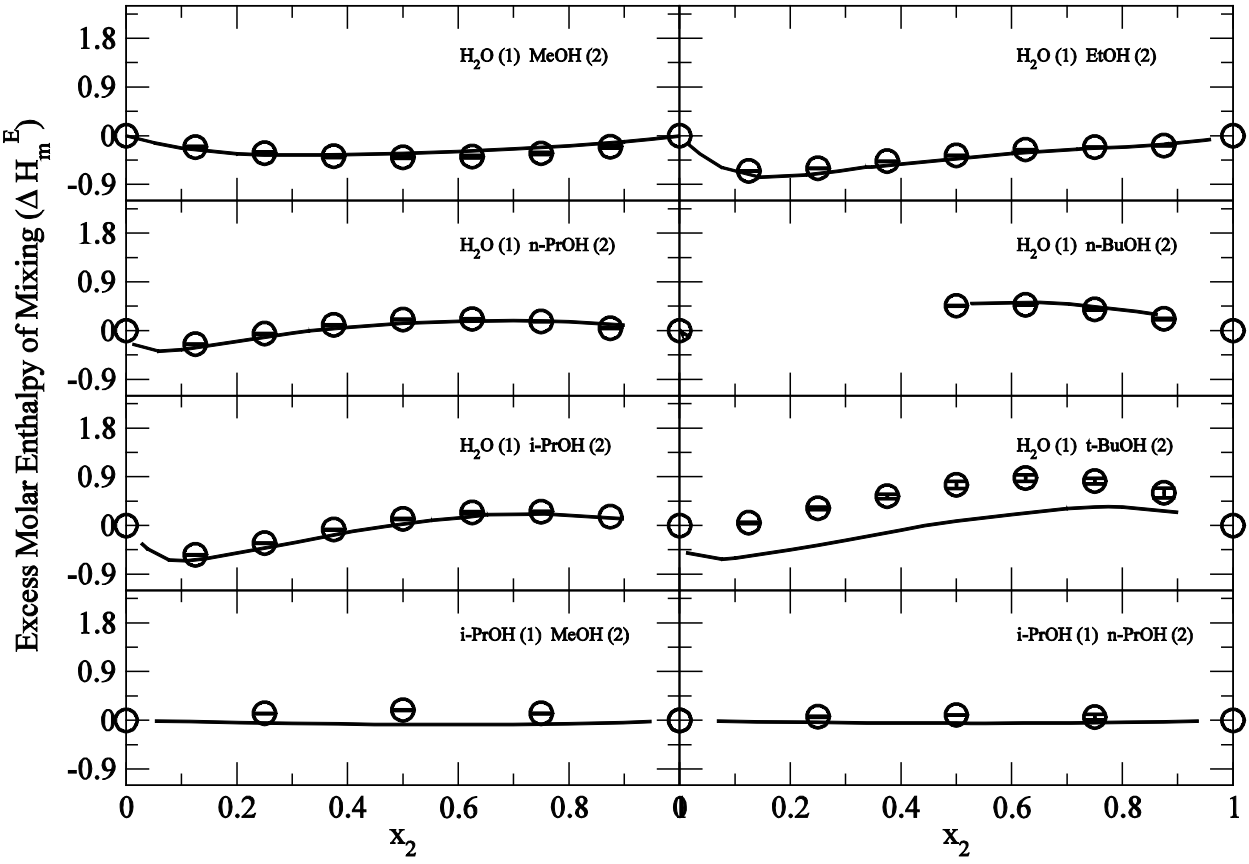
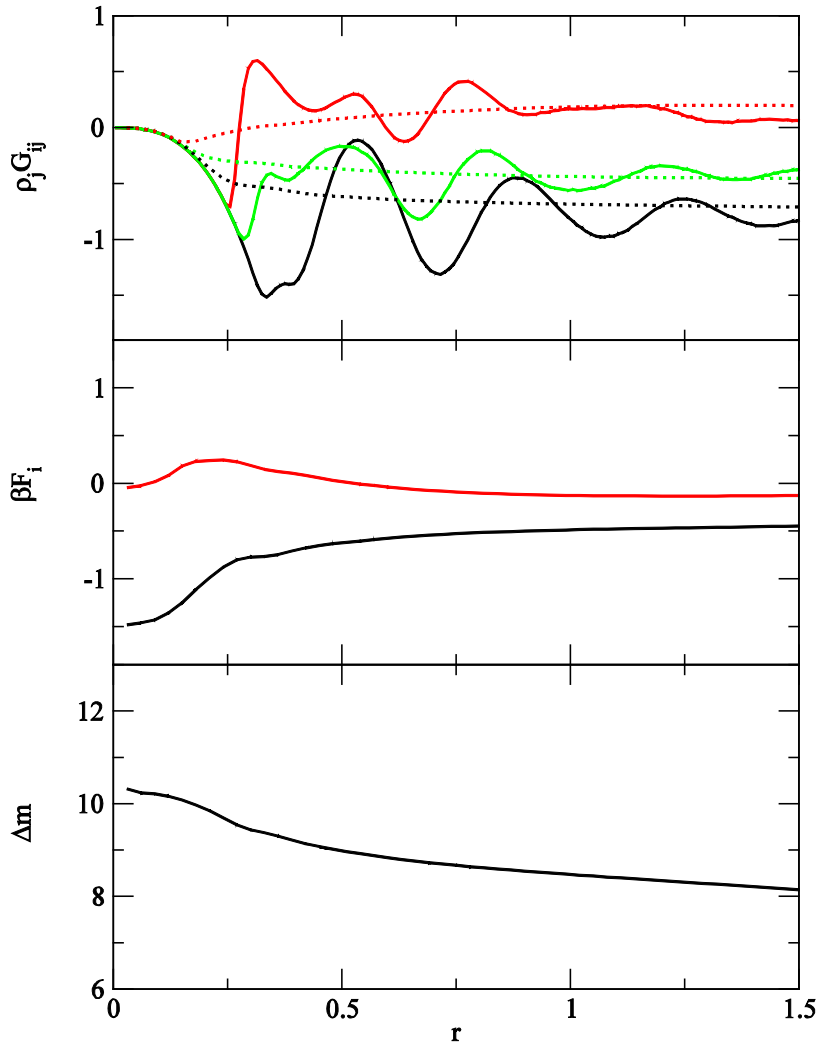


Figure 2.5 The local particle-particle (N_{ij}), particle-energy (F_i) and excess energy-energy fluctuations (Δ_m) as a function of the local sphere radius (r in nm) for a methanol and water mixture at a mole fraction of 0.5. Top: Excess coordination numbers N_{22} (black), N_{11} (red), and N_{21} (green). Solid lines correspond to integrals over the corresponding rdfs (Equation 2.22), while dashed lines correspond to values obtained from the number fluctuations (Equation 2.11). Middle: Particle-energy fluctuations for methanol (black) and water (red) (Equation 2.12). Bottom: Local excess energy fluctuations of solution (Equation 2.13).



Before examining the particle-energy and (excess) energy-energy fluctuations for the alcohol mixtures we determined the excess enthalpy of mixing for each solution. Clearly, if the models provide poor descriptions of the enthalpy of mixing they will also provide poor values for the excess partial molar enthalpies and, according to Equation 2.7, the local fluctuations cannot be in complete agreement with experiment. The excess enthalpies of mixing are provided in Figure 2.4 for each system. The agreement with experiment was excellent for all but one system. The results for t-BuOH and water mixtures were consistently more unfavorable than the experimental data. Unfortunately, none of the charge distributions tried here provided reasonable excess enthalpy of mixing. This problem was also encountered in a previous parameterization of t-BuOH.[23] A much more sensitive test of the force field models is provided by the excess partial molar enthalpies. However, as they are closely related to the particle-energy fluctuations we will focus on the fluctuations themselves.

Local particle-energy fluctuations have been determined previously from computer simulations, but not for systems or quantities that can be directly compared to experimental data. The question that immediately arises is - how large or small is the local volume? From the simulation point of view one requires a local volume that is large enough that it fully captures the local behavior, but at the same time small enough that the local region can be considered small in comparison with the remain bulk solution which then acts as a constant temperature and chemical potential bath. This will undoubtedly be system dependent. One way of investigating the behavior of the system is to plot the fluctuating quantities as a function of local volume. This is illustrated in Figure 2.5. As mentioned previously, the fluctuations display less structure than the equivalent quantities determined from the rdfs. However, their limiting values are essentially the same. All quantities appear to converge to a reasonable constant value between 1.0-1.5 nm. The convergence of the integrals over the rdfs is less obvious as some oscillation in the integrals is still evident. Fortunately, our approach of averaging these quantities over a solvation shell entered around 1.5 nm provides excellent agreement with the equivalent fluctuating quantities.

The simulated and experimentally extracted particle-energy fluctuations are provided in Figure 2.6. The experimental data display some interesting features. First, in the MeOH/HOH mixtures the F's essentially mimic the excess partial molar enthalpies. This is a consequence of the relatively small values of the KBIs which itself indicates a similar general affinity between solute-solute, solute-solvent and solvent-solvent pairs. This changes as the alkyl chain length

increases (as expected) and hence the local fluctuations for n-PrOH/HOH do not resemble the excess partial molar quantities. Here, the large values of the excess coordination numbers in the vicinity of x_2 = 0.2 not only indicate large fluctuations in the number of solute and solvent molecules, but also an accompanying large fluctuation in the local excess energy. The major trends in the experimental data were well reproduced by the simulations. Again, this provides added confidence that the current models are adequately reproducing the main thermodynamic properties of these solutions.

Figure 2.6 The local particle-energy fluctuations (F_i) as a function of alcohol mole fraction. The solid lines correspond to experimental data and the circles to the raw simulation data. Red lines correspond to the solvent (1) and black line to the solute (2).

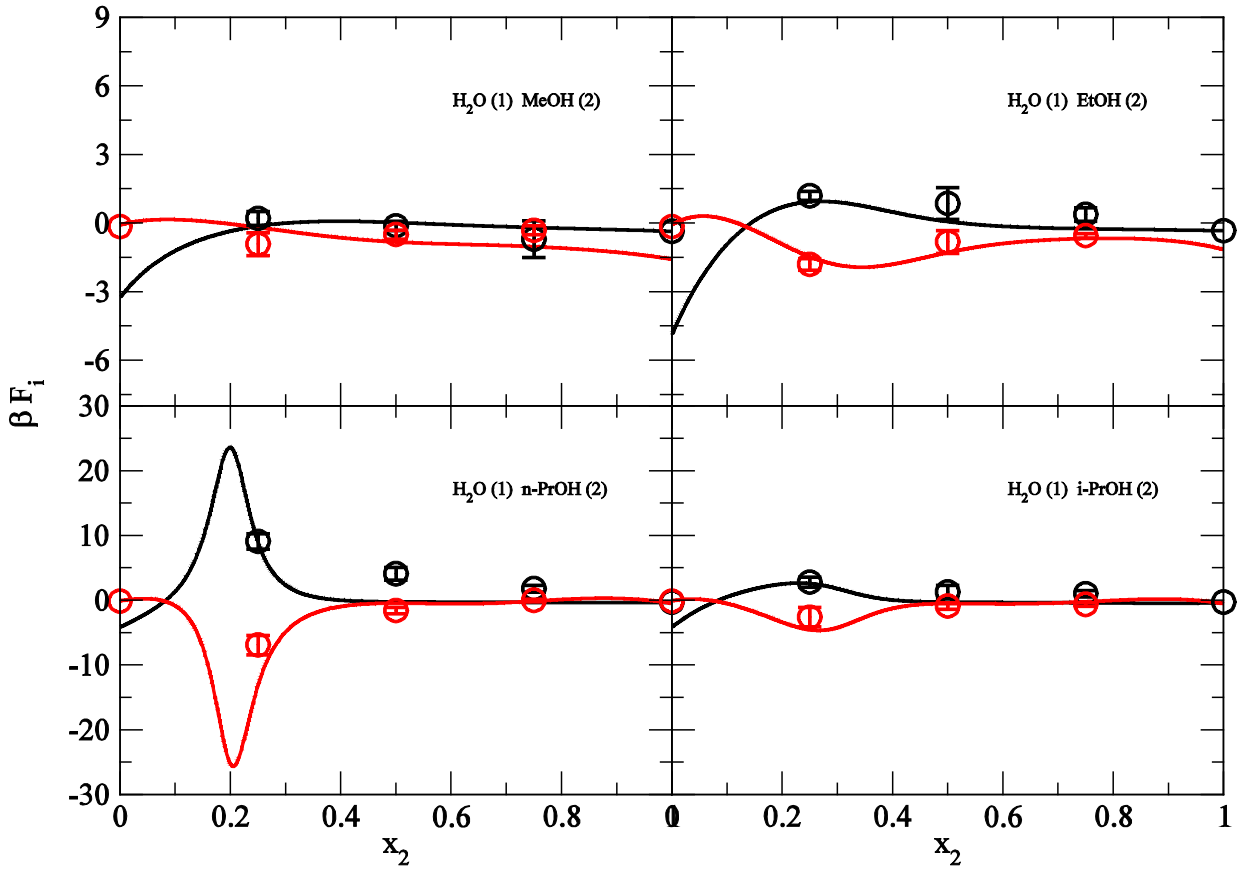
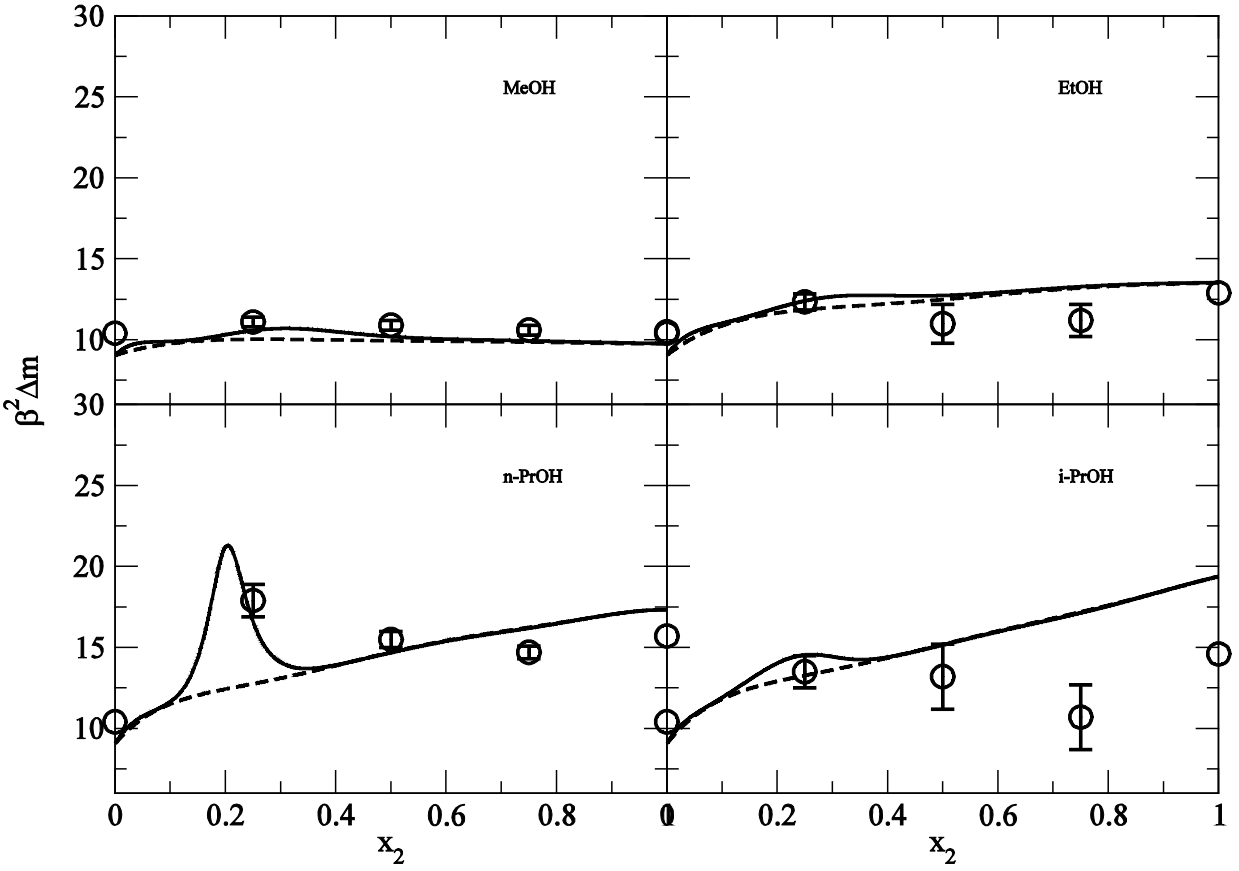


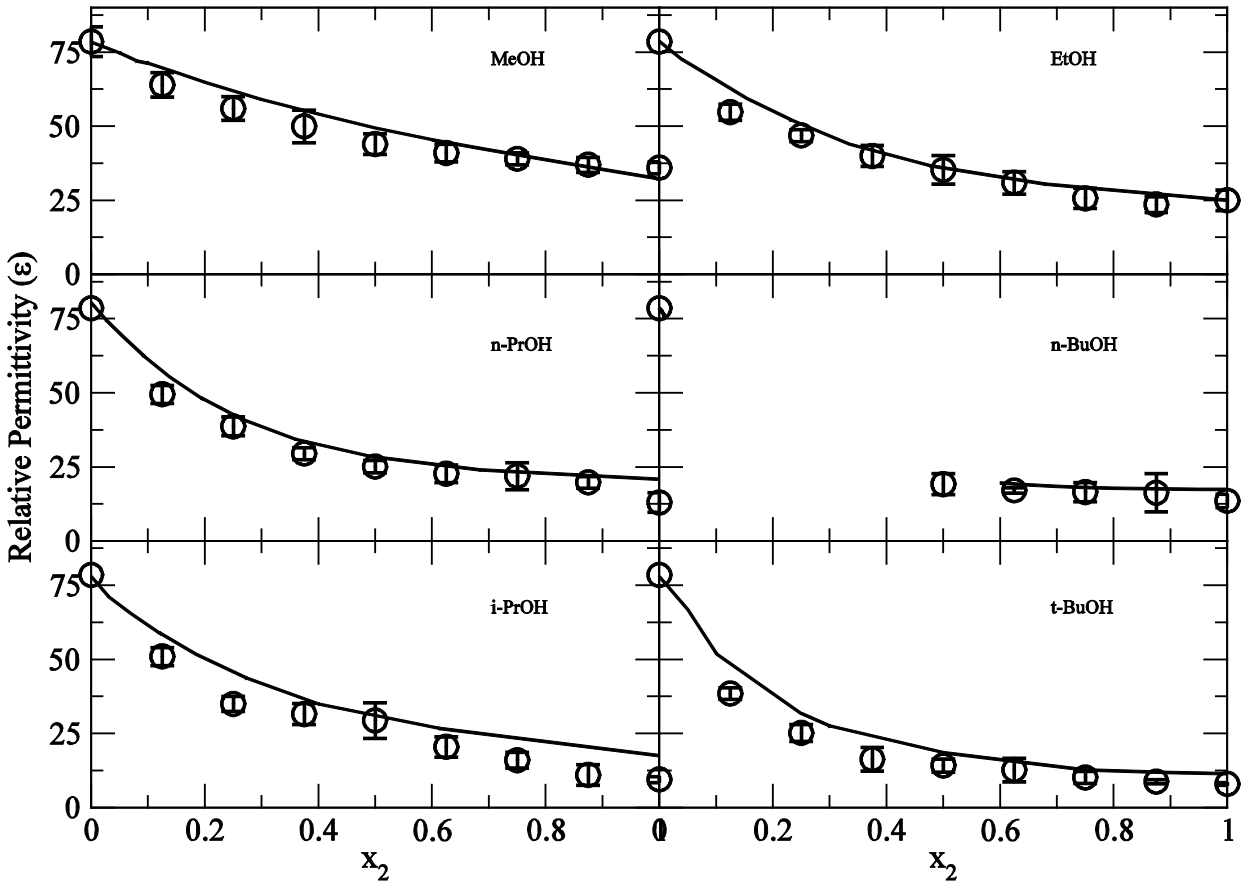
Figure 2.7 The excess energy-energy fluctuations (Δ_m) as a function of alcohol mole fraction. The solid lines correspond to the experimental data and the circles to the simulation data. The dashed lines correspond to the experimental data of C_p.



A similar pattern is observed for the energy-energy fluctuations displayed in Figure 2.7. The experimental excess energy fluctuation data mimic the heat capacity data for the small linear alcohols. However, n-PrOH mixtures with water indicate a large increase in the magnitude of the energy fluctuations in the vicinity of $x_2 = 0.2$. This feature appears to be reasonable well reproduced by the simulations. When taken together Figures 2.5-2.7 strongly suggest that the current models reproduce, almost quantitatively, the nontrivial composition dependence of the

particle and energy fluctuations observed for these systems. This in turn suggest that a wide range of thermodynamic properties are also reasonable accurate.

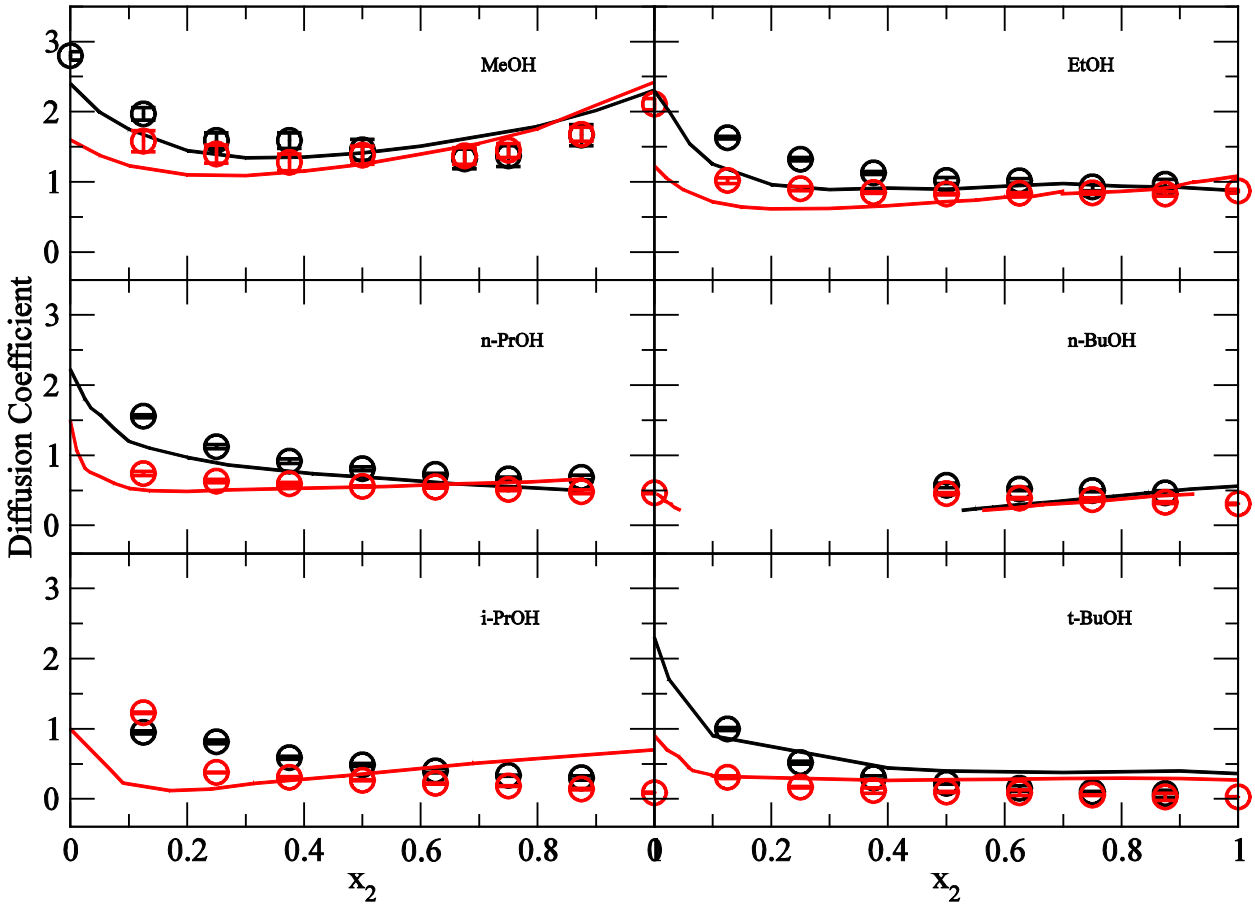
Figure 2.8 The relative permittivity as a function of the alcohol mole fraction. The solid lines correspond to the experimental data,[4,67-69] while the circles correspond to the simulation data.



An additional advantage of the KBFF approach is that one doesn't have to sacrifice agreement for other solution properties in an effort to accurately reproduce the local fluctuations. Here, we provide examples in the form of the relative permittivities of the solutions, and the translational diffusion constants for both the solutes and solvents. The relative permittivities are

displayed in Figure 2.8. The relative permittivities all change smoothly from the large value for pure water, to the lower values for the pure alcohols. This trend in the data is correctly reproduced by the simulations. In general, the agreement with experiment is very reasonable. A small underestimation of the permittivities can be traced to the slightly low value for pure SPC/E water.

Figure 2.9 The water (black) and alcohol (red) translational self-diffusion constants ($\times 10^{-9}$ $m^2 s^{-1}$) as a function of alcohol mole fraction. The solid lines correspond to the experimental data,[6,11,70-74] and the circles to the simulation data.



The self-diffusion coefficients of both water and alcohol are displayed in Figure 2.9. The experimental trends were reasonably well reproduced for MeOH, EtOH and PrOH. However, the results for n-BuOH and i-PrOH mixtures with water were relatively poor with both simulated diffusion constants decreasing with alcohol content compared to the increase exhibited by experiment. Some disagreement is to be expected considering the diffusion constant for pure SPC/E water is higher than experiment, while the simulated values for the pure alcohols are generally lower than experiment. Nevertheless, it is disappointing that these trends are not reproduced. Furthermore, it did not appear that other charge distributions provided any improvement in this area.

Table 2.2 Properties of pure liquid alcohols at 300 K compared with experimental data.[75-84]

		ρ (g/cm^3)	E_{pot} (kJ/mol)	ΔH_{vap} (kJ/mol)	D_c (10^{-9} m^2s^{-1})	κ_T (10^{-5} /bar)	ε	C_p (J/K/mol)	α_p (10^{-4} /K)
MeOH	md	0.742	-41.56	45.6	2.0	8	36	87	12
	exp	0.792		37.4	2.3	12	33	81	12
EtOH	md	0.755	-41.47	49.2	0.9	9	25	107	11
	exp	0.789		42.4	1.0	11	25	113	11
n-PrOH	md	0.771	-40.29	53.8	0.4	8	14	131	9
	exp	0.803		46.3	0.6	9	21	144	11
i-PrOH	md	0.741	-48.19	59.8	0.1	8	10	121	9
	exp	0.785		45.2	0.6	11	18	161	11
n-BuOH	md	0.781	-38.06	58.7	0.3	7	13	156	9
	exp	0.81		51.7	0.5	9	18	174	9
t-BuOH	md	0.778	-33.37	48.7	0.2	9	8	170	12
	exp	0.781		50.8	0.4	10	12	210	13
n-Octanol	md	0.808	-30.40	81.5	0.1	6	4	264	7
	exp	0.824		70.1	0.2	7	10	304	8

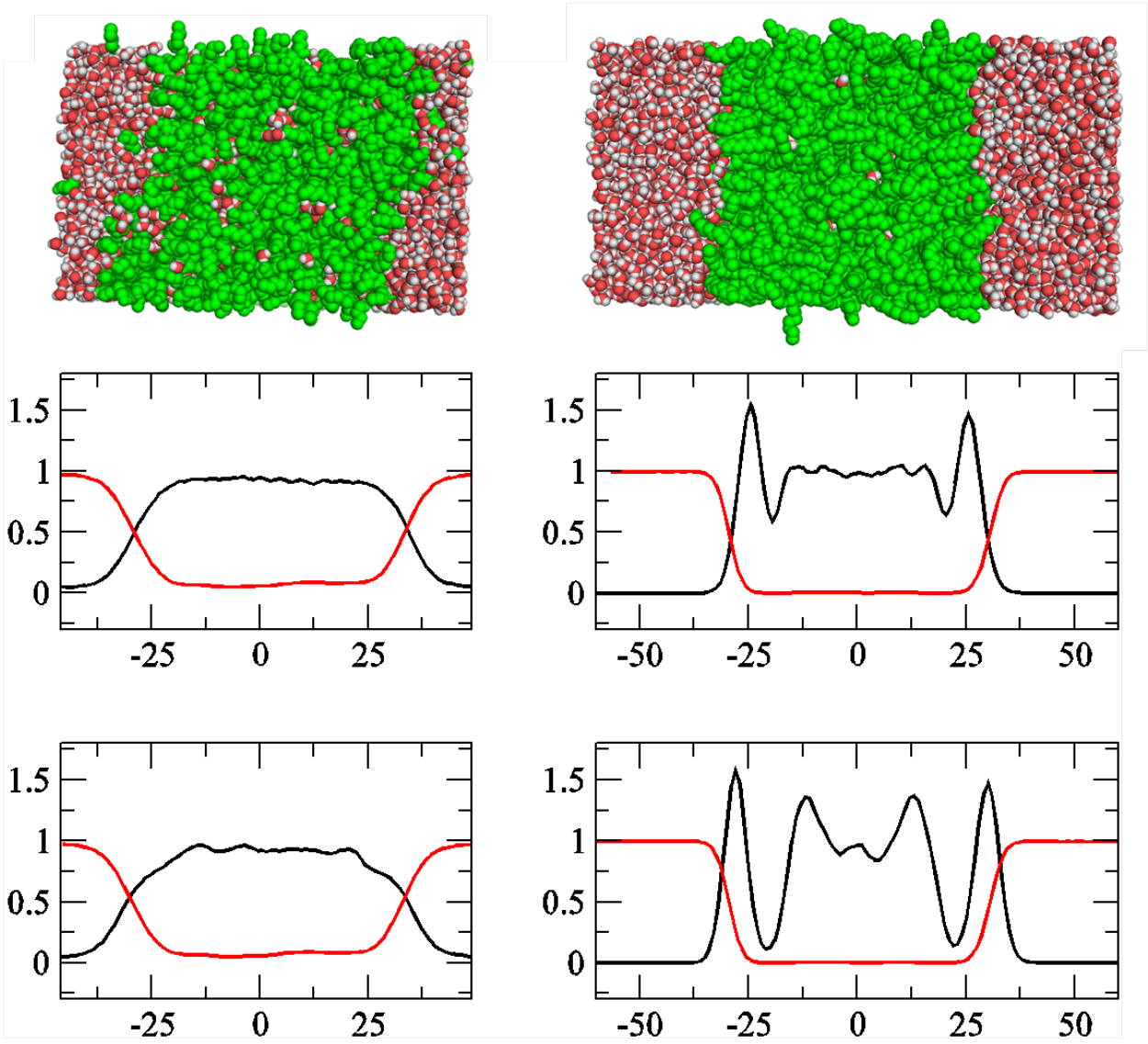
A selection of properties for the pure liquid alcohols is provided in Table 2.2. Most of the properties of pure alcohols can be reproduced reasonably accurately. The deviation of simulated

heat of vaporization is largest compared with other properties, especially when the carbon chain is longer, such as n-octanol.

In the current simulations we have used Kirkwood-Buff Theory to essentially analyze the atomic structure alcohol and water solutions. Experimentally, diffraction is commonly used to determine the atomic structure in solution. Several studies of small-angle X-ray scattering of alcohol-water mixtures have appeared, from which one can get information like atom-atom radial distribution function, number of hydrogen bond, isothermal compressibility and fluctuation numbers.[85-88] Moreover, some measurements were combined with Kirkwood-Buff analysis to exam heterogeneity of alcohols.[13] All properties mentioned above can be derived from simulation result as well. The comparison between experimental data and simulation result would be a good way to validate the accuracy of a force field.

One of the features of the linear alcohols is the change from fully miscible to immiscible as the length of the hydrocarbon chain increases. This can represent a stern test for new models and many examples of incorrect phase separated systems can be found. As a test of the current models we have performed simulations of n-BuOH and water at an alcohol mole fraction of 0.25, which is in the immiscible region. In addition, we have also simulated n-OctOH and water at an alcohol mole fraction of 0.10 to ensure demixing occurs in agreement with experiment. Both systems were simulated by randomly placing all molecules in an elongated box of 6×6×9 nm for n-BuOH and 6×6×12 nm for n-OctOH using anisotropic pressure coupling. The systems were then simulated for a total of 50 ns. In both simulations an interface starts to form within a few nanoseconds and eventually leads to a phase separation for both cases. Snapshots of both systems after 50 ns of simulation are shown in Figure 2.10. The corresponding density profile for both systems based on either the center of mass or oxygen positions, are also included in Figure 2.10, and are similar to previous simulation studies.

Figure 2.10 Snapshots and density profiles for the n-BuOH and water system (left), n-OctOH and water system (right). Top: snapshots of alcohol and water phase separation after 50 ns. Middle: center of mass density profile. Bottom: oxygen atom profile. Black lines indicate alcohol and red lines indicate water densities.



One clearly observes ordered n-octanol molecules at the interface where the hydrophilic hydroxyl group is in contact with the water phase leaving the long aliphatic chain in the central n-octanol rich region. This ordering is driven by hydrogen bonds between the interfacial water and n-octanol molecules. This provides significant structure to the oxygen density profile of n-octanol at the interface. Subsequently, there is then a significant depletion region just within the n-octanol side of the interface, indicating the presence of long hydrocarbon chain in hydrophobic layers. This is the same picture as observed from experiment by Steel et al.[89-91] The oxygen density then oscillates in the central organic region – again in agreement with previous MD simulations.[7,92] The snapshots and profiles clearly resemble those expected for a simple bilayer.

We could reproduce the essential properties of an n-butanol and water phase separation when starting from two components randomly mixed together. This is in consistent with the experimental fact that there is a miscibility gap in n-butanal/water mixtures (0.02 – 0.49 mole fraction of n-butanol) at 293K to 403K.[93,8] The saturated mole fraction of n-butanol was observed to be 0.018 in the water rich region of our simulations.

Simulations of n-octanol and water mixtures also phase separate as expected. The water molecules interact with n-octanol hydroxyl group. Most stay in the water rich region, although a small fraction of them remain in the interior of the bulk n-octanol phase. Studies on structure and solvation of both dry and wet n-octanol have been investigated both experimentally and theoretically. The mole fraction of water in n-octanol rich region was found to be 0.09. This is lower than the experimental values of 0.20 – 0.29,[94] but similar to a previous simulation result of 0.12.[7,79] However, it is clear from Figure 2.10 that there is only a small central region that may truly be considered as representative of the bulk n-octanol phase. Much larger systems would be required to obtain reliable values. Hence, the simulated mole fractions can only be considered as rough estimates. In terms of wet octanol study, since the solubility of n-octanol in water is very low (about 4.5×10^{-8} mol/L),[95] our system size is not big enough to correctly predict its molar fraction. However, our simulation result referring to interface between n-octanol and water indicates in a wet octanol phase interface, in agreement with experimental octanol/water interfaces.[7,89,92]

2.6 Conclusions

We have described the development of a Kirkwood-Buff derived force field for alcohols which accurately reproduces the experimental Kirkwood-Buff (KB) integrals observed in solution mixtures. The general alcohol models employed here reasonable reproduced the experimental KB integrals for a series of aqueous alcohol solutions over a range of compositions.

The emphasis has been on the development of accurate charge distributions for the different alcohol models using a KB analysis of the experimental data as a guide. In our opinion, this is a demonstrated strength of the KB force field approach.[48,86] The final charge on the alcohol oxygen was -0.82 (-0.902 for the oxygen of i-propanol), which is significantly larger than that for many previous non-polarizable force fields.[20,52] It should be noted that a comparison of just effective charges from different force fields is difficult as one can produce similar effective interactions between different groups by varying the charge and size (σ) parameters accordingly. Our alcohol and water oxygen charges are very similar and probably should be as the electronegativities of carbon and hydrogen are very similar. This results in the correct balance between solvation of alcohol by other alcohol molecules and solvation by water molecules as displayed by the KB integrals. We presented a modified charge distribution for i-propanol in an effort to improve the properties for this alcohol.

Here we have studied the local properties of a series of alcohol and water solutions by determining the appropriate fluctuating quantities (KBIs, F's and Δm). In the majority of cases, the models quantitatively reproduce the thermodynamic properties of mixtures of water and alcohol up to alcohol mole fractions of 0.1. At higher mole fractions some small deviations are observed which appear to relate to deviations in the enthalpy of mixing and partial molar enthalpies of alcohol and water. It is possible that these small deviations could be due to environmentally dependent polarization effects, although this is difficult to determine.

The current models and results indicate that is a transferable and additive force field, with the exception of i-propanol. The results strongly suggest that models developed for small linear alcohols can be used with confidence for larger linear alcohols. The situation for secondary and tertiary alcohols is more complicated. This is to expected as the local environment around the alcohol molecules in solution will be very different.

2.7 Supporting Information

The following tables and figures provide more details about this study. Table 2.3 gives all bonded parameters of KB force field applied for simulation. Table 2.4 shows summary of simulation results including both pure liquid and mixtures. Table 2.5 compares the proper dihedral distribution in gas phase and condensed phase. Figure 2.11 and 2.12 give radial distribution functions (g_{11}, g_{12}, g_{22}) based on distance measured from center-of-mass and atom-atom for different compositions. In Figure 2.13, calculated excess coordination numbers (N_{11}, N_{12}, N_{22}) are in good agreement with experimental data. Density in Figure 2.14 can reproduce well for most of alcohols and their mixtures, though the pure propanol density is a little lower than expected. Charges on t-butanol was discussed, Figure 2.15 shows properties with extra 20% charge on oxygen, though the excess coordination number gives reasonable results, relative permittivity is very low at pure t-butanol, therefore we prefer to keep the original charge. In Figure 2.16, first and second shell coordination numbers are compared with different alcohols and it gives a decreased trend as alcohol carbon chain increased. Finally to validate that the 10% charge increase on i-propanol can reproduce physical properties better, methanol and i-propanol mixture systems were studied to test relative permittivity, diffusion coefficients as shown in Figure 2.17.

Table 2.3 Bonded force field parameters used in the simulations.

bonds	r_0		r_0
CH_n-O	0.1430	CH_n-CH_n	0.1530
O-H	0.0945		
Angles		k_θ	θ_0
MEOH	H-O-CH_n	450.0	109.5
ETOH	O-CH_2-CH_3	520.0	109.5
n-PROH/ BUOH	CH_n-CH_n-CH_n	530.0	111.0
i-PROH	O-CH_1-CH_3	530.0	111.0
t-BUOH	O-C-CH_3	610.0	108.0
	CH_3-C-CH_3	530.0	112.0
Dihedrals	k_φ	φ_0	*n*
CH_3-CH_n-O-H	0.85	0.0	1
	0.40	0.0	2
	3.00	0.0	3
CH_n-CH_2-CH_2-O	2.55	0.0	1
	1.20	0.0	2
	9.00	0.0	3
CH_3-CH_2-CH_2-CH_2	5.00	0.0	1
	2.00	0.0	2
	7.00	0.0	3
Impropers	k_ξ		ξ_o
CH_1-O-CH_3-CH_3	334.8		35.26

Table 2.4 Summary of the Alcohol and Water Simulations.[a]

x_c	N_c	V, nm^3	ρ_c, M	ρ, g/cm^3	E_{pot}, kJ/mol	T_{sim}, ns
H_2O						
0.0	2000	60.16	0.0	0.995	-46.45	3
Methanol/ H_2O						
0.125	779	213.334	6.06	0.953	-46.48	5
0.25	1374	212.024	10.76	0.916	-46.13	5
0.375	1843	212.144	14.43	0.881	-45.65	5
0.50	2223	213.315	17.30	0.849	-45.07	5
0.625	2536	215.142	19.57	0.819	-44.40	5
0.75	2799	217.595	21.36	0.791	-43.62	5
0.875	3022	220.408	22.77	0.765	-42.68	5
1.0	3215	223.327	23.91	0.742	-41.56	5
Ethanol/H_2O						
0.125	702	211.834	5.50	0.948	-46.29	30
0.25	1152	212.379	9.01	0.902	-45.61	30
0.375	1465	214.165	11.36	0.864	-44.89	30
0.50	1695	216.093	13.03	0.835	-44.20	30
0.625	1871	217.894	14.26	0.811	-43.55	30
0.75	2010	219.886	15.18	0.790	-42.92	30
0.875	2123	222.022	15.88	0.772	-42.27	30
1.0	2216	224.542	16.39	0.755	-41.47	30
n-Propanol/ H_2O						
0.125	646	214.884	4.99	0.930	-45.78	30
0.25	1006	216.577	7.71	0.881	-44.85	30
0.375	1236	217.981	9.42	0.849	-43.96	30
0.50	1395	219.099	10.57	0.826	-43.12	30
0.625	1512	220.151	11.40	0.809	-43.36	30
0.75	1602	221.279	12.02	0.795	-41.66	30
0.875	1672	222.302	12.49	0.783	-41.01	30
1.0	1730	223.816	12.84	0.771	-40.29	30

n-Butanol/ H_2O						
0.50	1188	221.111	8.92	0.822	-41.82	30
0.625	1271	221.697	9.52	0.809	-40.77	30
0.75	1334	222.374	9.96	0.798	-39.80	30
0.875	1383	223.095	10.29	0.789	-38.93	30
1.0	1422	223.974	10.54	0.781	-38.06	30
i-Propanol/ H_2O						
0.125	642	213.674	4.99	0.929	-47.04	30
0.25	996	216.536	7.64	0.872	-47.04	30
0.375	1221	219.358	9.24	0.899	-47.03	30
0.50	1376	221.611	10.31	0.805	-47.07	30
0.625	1490	223.633	11.06	0.784	-47.16	30
0.75	1576	225.115	11.63	0.768	-47.37	30
0.875	1645	226.884	12.04	0.755	-47.67	30
1.0	1700	228.505	12.36	0.742	-48.19	30
t-Butanol/ H_2O						
0.125	588	210.696	4.63	0.928	-47.74	30
0.25	873	212.296	6.83	0.875	-48.41	30
0.375	1041	214.045	8.08	0.841	-49.06	30
0.50	1151	215.137	8.89	0.820	-49.78	30
0.625	1230	216.343	9.44	0.802	-50.45	30
0.75	1288	217.059	9.86	0.790	-51.20	30
0.875	1334	217.893	10.17	0.780	-52.05	30
1.0	1370	216.664	10.50	0.778	-53.48	30
Methanol/i-Propanol						
0.25	482	228.052	3.51	0.741	-46.36	10
0.50	1112	227.181	8.13	0.741.	-44.65	10
0.75	1976	225.921	14.52	0.741	-43.07	10
Propanol/i-Propanol						
0.25	427	227.428	3.12	0.749	-46.11	10
0.50	857	226.176	6.29	0.756	-44.11	10

0.75	1292	225.198	9.53	0.764	-42.17	10

[a] All simulations were performed at 300 K and 1 atm in the NpT ensemble. E_{pot} is the average potential energy per molecule and T_{sim} is the total simulation time. All other symbols have their usual meaning (see text).

Table 2.5 Summary of dihedral distributions.

	CCOH			CCCO			CCCC		
	-60	60	180	-60	60	180	-60	60	180
gas	30%	30%	40%	34%	34%	32%	14%	14%	72%
liquid	30%	30%	40%	23%	23%	54%	10%	12%	78%

Figure 2.11 Center of mass radial distribution functions as a function of distance (nm). Mole fractions of 0.25 (black), 0.50 (red), and 0.75 (green) are displayed.

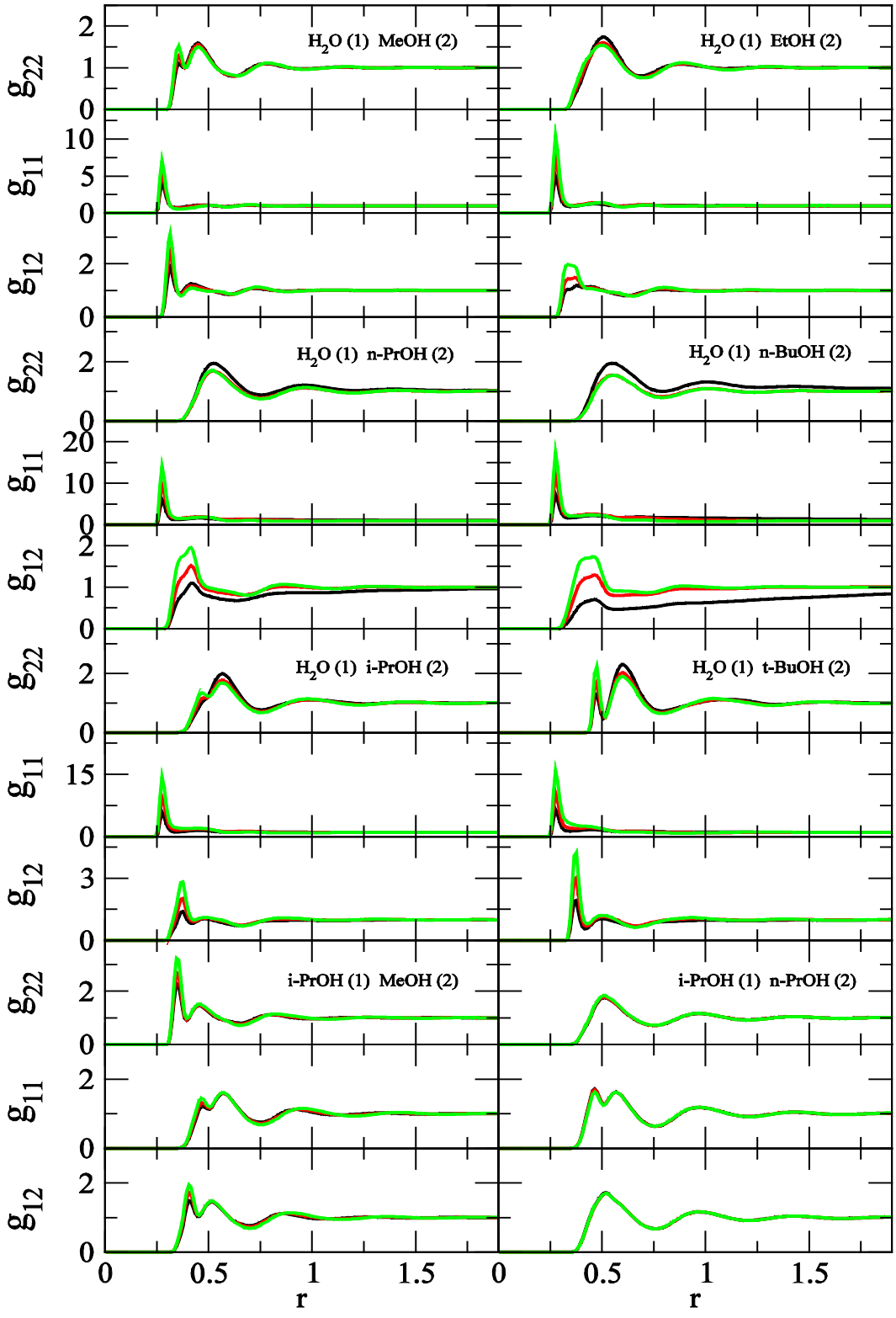


Figure 2.12 Oxygen-oxygen radial distribution functions as a function of distance (nm). Mole fractions of 0.25 (black), 0.50 (red), and 0.75 (green) are displayed.

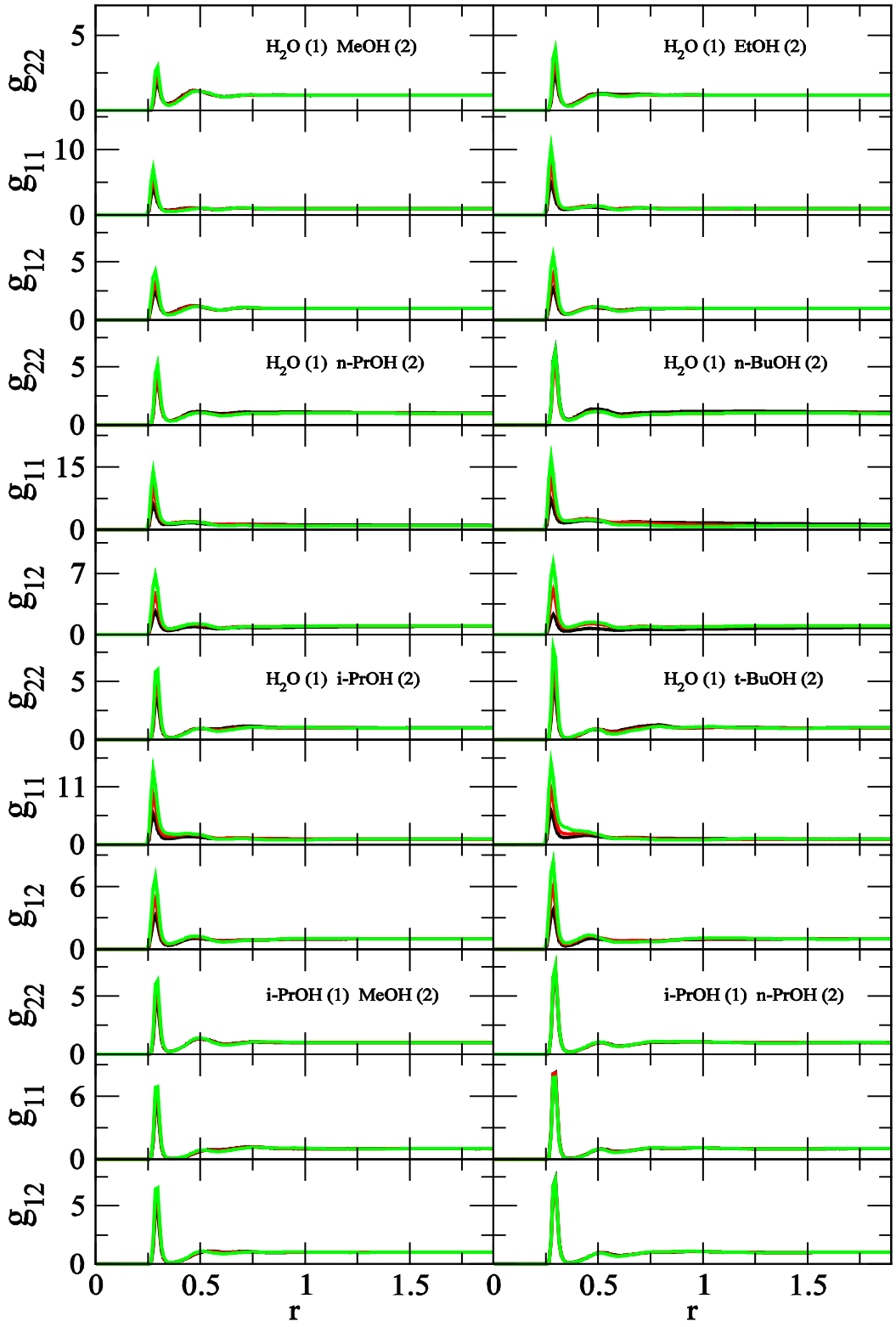


Figure 2.13 Excess coordination numbers (N_{ij}) as a function of alcohol mole fraction. The solid lines correspond to experimental data, the circles to raw simulation data.

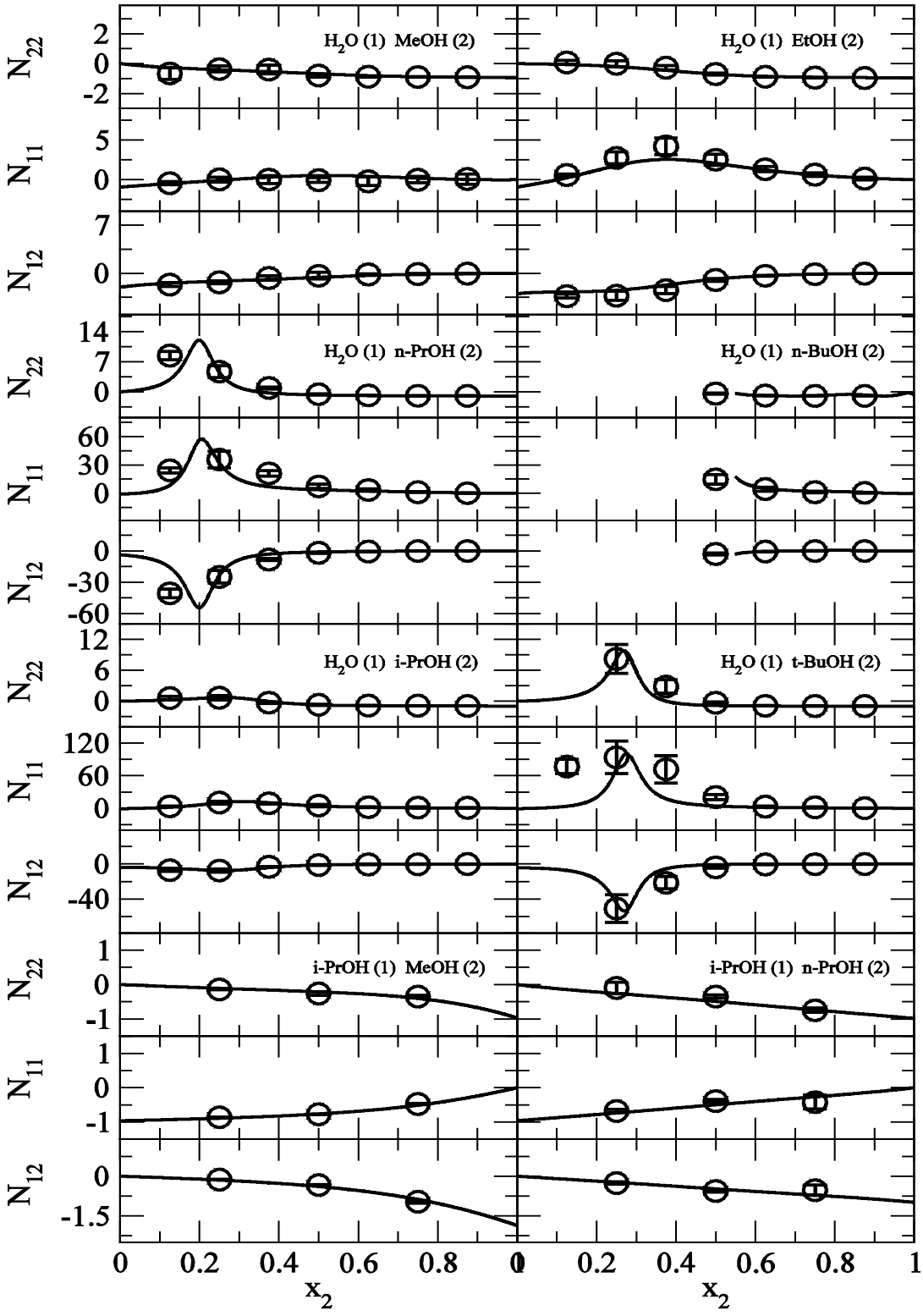


Figure 2.14 The density (g/cm^3) as a function of the alcohol mole fraction. The solid lines correspond to the experimental data, the circles to the raw simulation data.

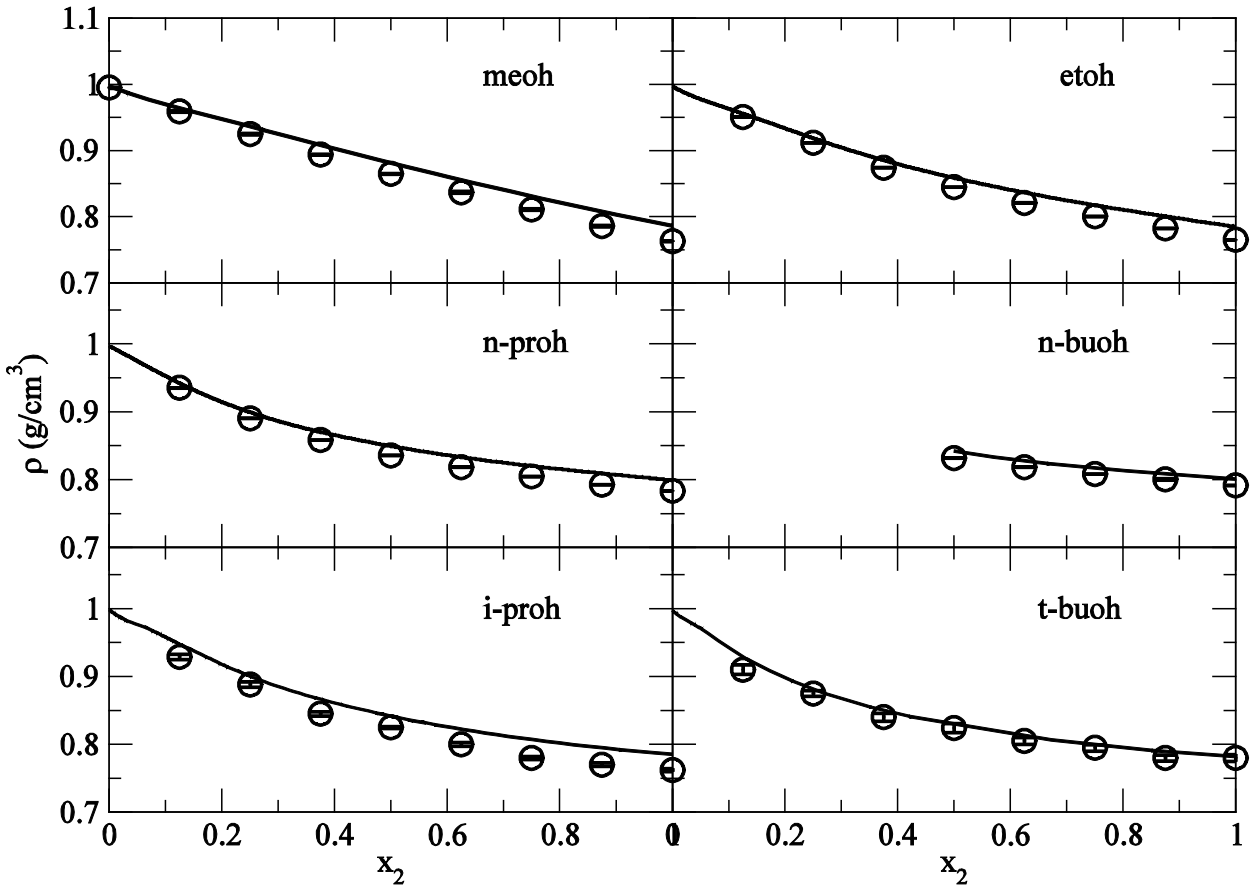
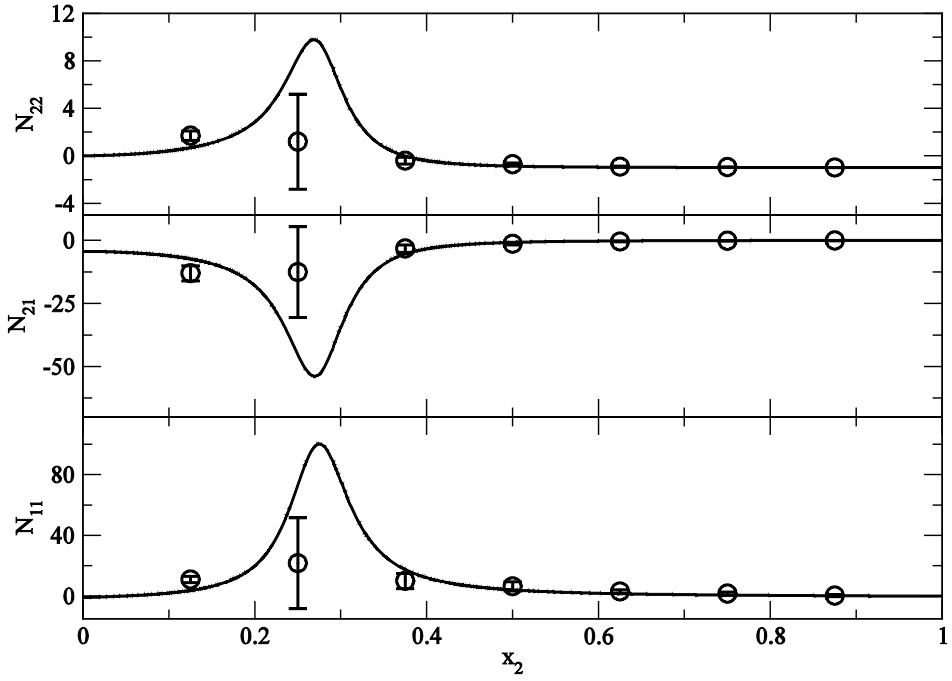


Figure 2.15 Excess coordination numbers and relative permittivity of t-butanol with 20% more charge.



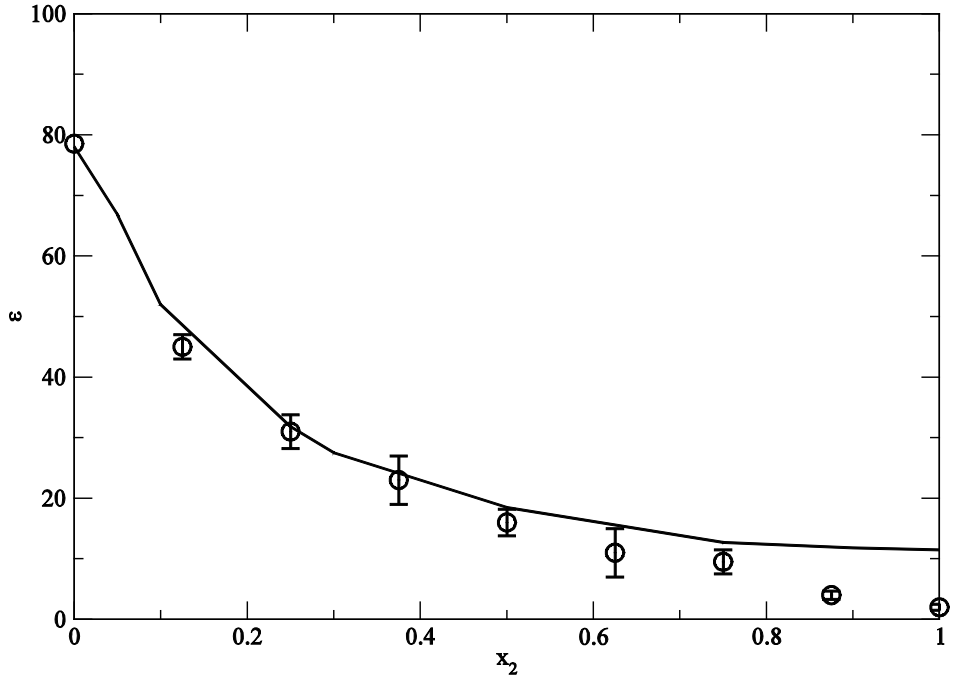


Figure 2.16 Alcohol to water oxygen-oxygen coordination numbers (n_{21}) for the first and second hydration shells as a function of alcohol mole fraction for methanol (black), ethanol (red) and n-propanol (green) and n-butanol (blue) are displayed.

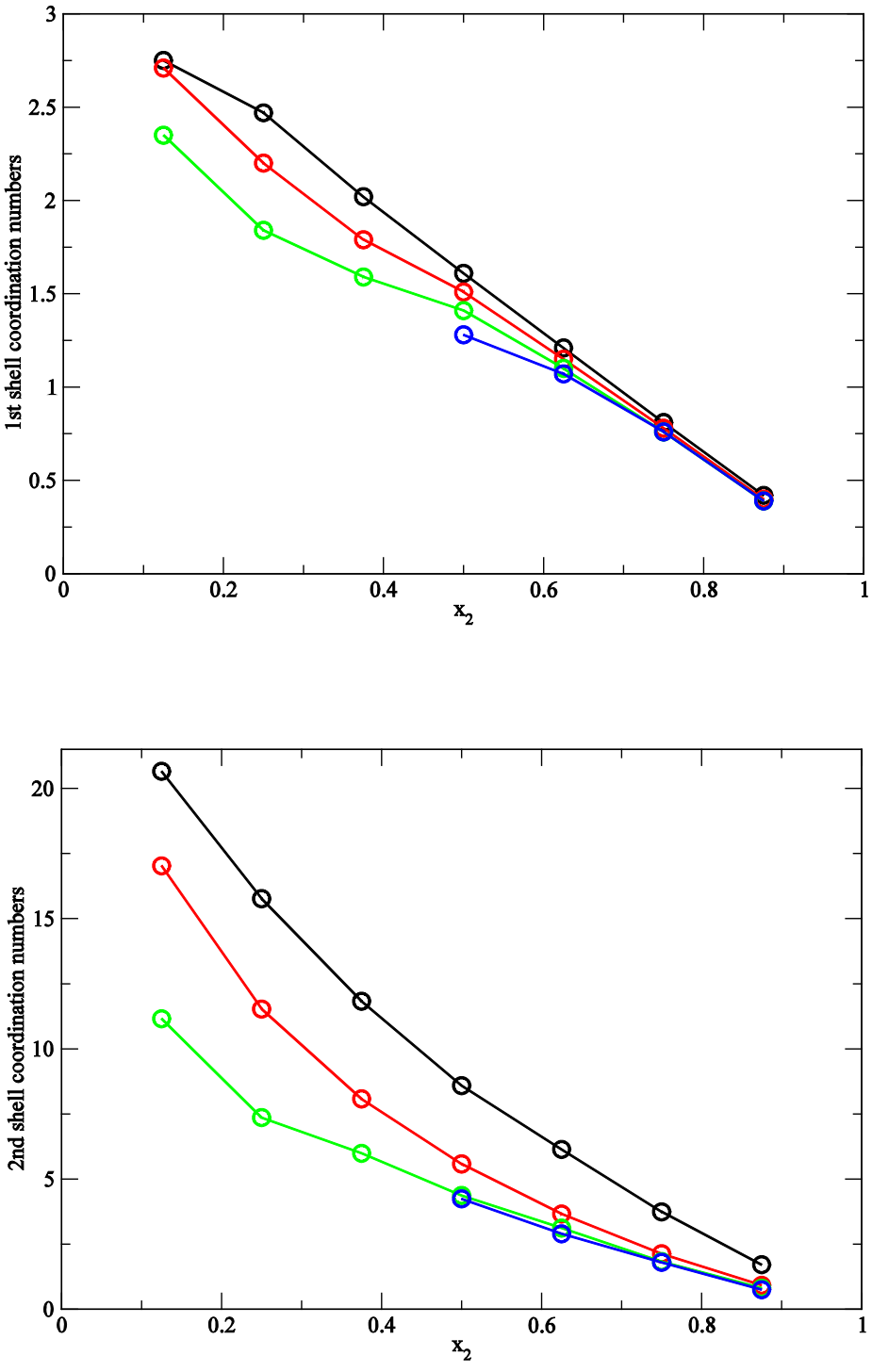
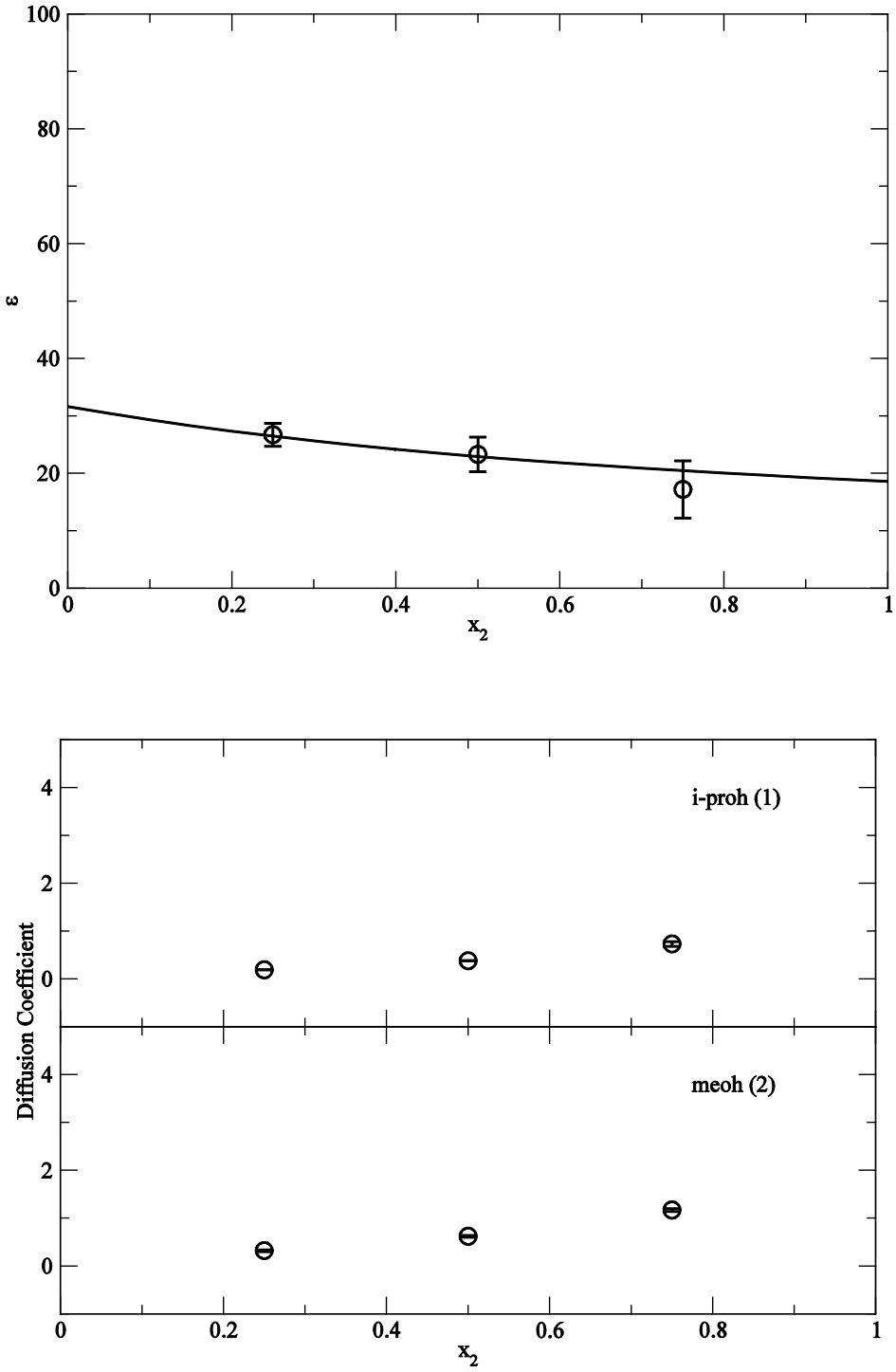


Figure 2.17 Relative permittivity and diffusion coefficient as a function of i-propanol mole fraction in methanol solvent.[60,96]



2.9 References

(1) Benjamin, L.; Benson, G. C. *Journal of Physical Chemistry* **1963**, *67*, 858.

(2) Benson, G. C.; Kiyohara, O. *Journal of Solution Chemistry* **1980**, *9*, 791.

(3) Lama, R. F.; Lu, B. C. Y. *Journal of Chemical and Engineering Data* **1965**, *10*, 216.

(4) Sengwa, R. J.; Madhvi; Sankhla, S. *Physics and Chemistry of Liquids* **2006**, *44*, 637.

(5) Tanaka, H.; Nakanishi, K.; Touhara, H. *Journal of Chemical Physics* **1984**, *81*, 4065.

(6) Wensink, E. J. W.; Hoffmann, A. C.; van Maaren, P. J.; van der Spoel, D. *Journal of Chemical Physics* **2003**, *119*, 7308.

(7) Napoleon, R. L.; Moore, P. B. *Journal of Physical Chemistry B* **2006**, *110*, 3666.

(8) Arlt, W.; Macedo, M. E. A.; Rasmussen, P.; Soerensen, J. M. *Liquid-Liquid Equilibrium Data Collection. Part 1: Binary Systems* **1979**; Vol. V.

(9) Boyne, J. A.; Williams.Ag *Journal of Chemical and Engineering Data* **1967**, *12*, 318.

(10) Chitra, R.; Smith, P. E. *Journal of Physical Chemistry B* **2000**, *104*, 5854.

(11) Derlacki, Z. J.; Easteal, A. J.; Edge, A. V. J.; Woolf, L. A.; Roksandic, Z. *Journal of Physical Chemistry* **1985**, *89*, 5318.

(12) Koga, Y.; Westh, P.; Nishikawa, K. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **2003**, *81*, 150.

(13) Nishikawa, K.; Iijima, T. *Journal of Physical Chemistry* **1993**, *97*, 10824.

(14) Biscay, F.; Ghoufi, A.; Lachet, V.; Malfreyt, P. *Journal of Physical Chemistry C* **2011**, *115*, 8670.

(15) Biscay, F.; Ghoufi, A.; Malfreyt, P. *Journal of Chemical Physics* **2011**, *134*.

(16) Nakanishi, K.; Ikari, K.; Okazaki, S.; Touhara, H. *Journal of Chemical Physics* **1984**, *80*, 1656.

(17) Moon, S. D.; Kang, Y. S.; Lee, D. J. *Macromolecular Research* **2007**, *15*, 491.

(18) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *Journal of Physical Chemistry B* **1998**, *102*, 3586.

(19) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179.

(20) Jorgensen, W. L. *Journal of Physical Chemistry* **1986**, *90*, 1276.

(21) Walser, R.; Mark, A. E.; van Gunsteren, W. F.; Lauterbach, M.; Wipff, G. *Journal of Chemical Physics* **2000**, *112*, 10450.

(22) Rai, N.; Wagner, A. J.; Ross, R. B.; Siepmann, J. I. *Journal of Chemical Theory and Computation* **2008**, *4*, 136.

(23) Lee, M. E.; van der Vegt, N. F. A. *Journal of Chemical Physics* **2005**, *122*, 114509.

(24) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2005**, *109*, 15080.

(25) Gao, J. L.; Habibollazadeh, D.; Shao, L. *Journal of Physical Chemistry* **1995**, *99*, 16460.

(26) Gonzalez, M. A.; Enciso, E.; Bermejo, F. J.; Bee, M. *Journal of Chemical Physics* **1999**, *110*, 8045.

(27) Caldwell, J. W.; Kollman, P. A. *Journal of Physical Chemistry* **1995**, *99*, 6208.

(28) Noskov, S. Y.; Lamoureux, G.; Roux, B. *Journal of Physical Chemistry B* **2005**, *109*, 6705.

(29) Patel, S.; Brooks, C. L. *Journal of Chemical Physics* **2005**, *123*, 164502.

(30) Patel, S.; Brooks, C. L. *Journal of Chemical Physics* **2005**, *122*, 024508.

(31) Zhong, Y.; Patel, S. *Journal of Physical Chemistry B* **2010**, *114*, 11076.

(32) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. *Journal of Chemical Theory and Computation* **2007**, *3*, 1927.

(33) Cieplak, P.; Caldwell, J.; Kollman, P. *Journal of Computational Chemistry* **2001**, *22*, 1048.

(34) Paschek, D.; Geiger, A.; Herve, M. J.; Suter, D. *Journal of Chemical Physics* **2006**, *124*, 14508.

(35) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2003**, *107*, 3891.

(36) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2004**, *121*, 2180.

(37) Gee, M. B.; Smith, P. E. *Journal of Chemical Physics* **2009**, *131*.

(38) Bentenitis, N.; Cox, N. R.; Smith, P. E. *Journal of Physical Chemistry B* **2009**, *113*, 12306.

(39) Chen, F.; Smith, P. E. *Abstracts of Papers of the American Chemical Society* **2009**, *238*.

(40) Smith, P. E. *Journal of Chemical Physics* **2008**, *129*.

(41) Kang, M.; Smith, P. E. *Journal of Chemical Physics* **2008**, *128*.

(42) Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochemistry and Biophysics* **2008**, *50*, 1.

(43) Kang, M.; Smith, P. E. *Fluid Phase Equilibria* **2007**, *256*, 14.

(44) Kang, M.; Smith, P. E. *Journal of Computational Chemistry* **2006**, *27*, 1477.

(45) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, *119*, 11342.

(46) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, *118*, 10663.

(47) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *Journal of Physical Chemistry* **1987**, *91*, 6269.

(48) Kirkwood, J. G.; Buff, F. P. *Journal of Chemical Physics* **1951**, *19*, 774.

(49) Ben-Naim, A. *Statistical Thermodynamics for Chemists and Biochemists*; Plenum Press: New York, **1992**.

(50) Jiao, Y.; Smith, P. E. *Journal of Chemical Physics* **2011**, *135*, 014502.

(51) Ploetz, E. A.; Smith, P. E. *Journal of Chemical Physics* **2011**, *135*, 044506.

(52) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Journal of Computational Chemistry* **1998**, *19*, 535.

(53) *International Critical Tables VolIII pp120-121*

(54) Mitchell, A. G.; Wynnejones, W. F. K. *Discussions of the Faraday Society* **1953**, 161.

(55) Nakanish.K; Kato, N.; Maruyama, M. *Journal of Physical Chemistry* **1967**, *71*, 814.

(56) Redlich, O.; Kister, A. T. *Industrial and Engineering Chemistry* **1948**, *40*, 345.

(57) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684.

(58) Darden, T.; York, D.; Pedersen, L. *Journal of Chemical Physics* **1993**, *98*, 10089.

(59) Smith, P. E.; Vangunsteren, W. F. *Journal of Chemical Physics* **1994**, *100*, 3169.

(60) Chmielewska, A.; Zurada, M.; Klimaszewski, K.; Bald, A. *Journal of Chemical and Engineering Data* **2009**, *54*, 801.

(61) Hu, J. H.; Haynes, C. A.; Wu, A. H. Y.; Cheung, C. M. W.; Chen, M. M.; Yee, E. G. M.; Ichioka, T.; Nishikawa, K.; Westh, P.; Koga, Y. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **2003**, *81*, 141.

(62) Romero, C. M.; Paez, M. S. *Journal of Solution Chemistry* **2007**, *36*, 237.

(63) Liltorp, K.; Westh, P.; Koga, Y. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **2005**, *83*, 420.

(64) Tamura, K.; Hu, J. H.; Trandum, C.; Westh, P.; Haynes, C. A.; Koga, Y. *Physical Chemistry Chemical Physics* **2000**, *2*, 355.

(65) Franks, F. *Annals of the New York Academy of Sciences* **1965**, *125*, 277.

(66) Pratt, K. C.; Wakeham, W. A. *Proceedings of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences* **1975**, *342*, 401.

(67) Gregorowicz, J.; Bald, A.; Szejgis, A.; Chmielewska, A. *Journal of Molecular Liquids* **2000**, *84*, 149.

(68) Kaatze, U.; Schumacher, A.; Pottel, R. *Berichte Der Bunsen-Gesellschaft-Physical Chemistry Chemical Physics* **1991**, *95*, 585.

(69) Yilmaz, H.; Guler, S.; Guler, C. *Physica Scripta* **1999**, *59*, 77.

(70) Bulicka, J.; Prochazka, J. *Journal of Chemical and Engineering Data* **1976**, *21*, 452.

(71) Hawlicka, E.; Grabowski, R. *Journal of Physical Chemistry* **1992**, *96*, 1554.

(72) Ito, N.; Saito, K.; Kato, T.; Fujiyama, T. *Bulletin of the Chemical Society of Japan* **1981**, *54*, 991.

(73) Price, W. S.; Ide, H.; Arata, Y. *Journal of Physical Chemistry A* **2003**, *107*, 4784.

(74) Zhang, L.; Wang, Q.; Liu, Y. C.; Zhang, L. Z. *Journal of Chemical Physics* **2006**, *125*.

(75) Klofutar, C.; Nemec, T. *Journal of Solution Chemistry* **1996**, *25*, 1151.

(76) Marcus, Y.; Hefter, G. T. *Journal of Molecular Liquids* **1997**, *73-4*, 61.

(77) *Handbook of Tables for Applied Engineering Science*

(78) Zorebski, E.; Waligora, A. *Journal of Chemical and Engineering Data* **2008**, *53*, 591.

(79) Chen, B.; Siepmann, J. I. *Journal of the American Chemical Society* **2000**, *122*, 6464.

(80) Vuks, M. F. *Zhurnal Strukturnoi Khimii* **1973**, *14*, 730.

(81) Omelyan, I.; Kovalenko, A.; Hirata, F. *Journal of Theoretical & Computational Chemistry* **2003**, *2*, 193.

(82) Egorov, G. I.; Makarov, D. M. *Journal of Chemical Thermodynamics* **2011**, *43*, 430.

(83) Nakagawa, M.; Inubushi, H.; Moriyoshi, T. *Journal of Chemical Thermodynamics* **1981**, *13*, 171.

(84) Chickos, J. S.; Acree, W. E. *Journal of Physical and Chemical Reference Data* **2003**, *32*, 519.

(85) Guo, J. H.; Luo, Y.; Augustsson, A.; Kashtanov, S.; Rubensson, J. E.; Shuh, D. K.; Agren, H.; Nordgren, J. *Physical Review Letters* **2003**, *91*.

(86) Nishikawa, K. *Chemical Physics Letters* **1986**, *132*, 50.

(87) Nishikawa, K.; Kodera, Y.; Iijima, T. *Journal of Physical Chemistry* **1987**, *91*, 3694.

(88) Yamaguchi, T.; Benmore, C. J.; Soper, A. K. *Journal of Chemical Physics* **2000**, *112*, 8976.

(89) Steel, W. H.; Walker, R. A. *Nature* **2003**, *424*, 296.

(90) Steel, W. H.; Beildeck, C. L.; Walker, R. A. *Journal of Physical Chemistry B* **2004**, *108*, 16107.

(91) Steel, W. H.; Damkaci, F.; Nolan, R.; Walker, R. A. *Journal of the American Chemical Society* **2002**, *124*, 4824.

(92) Benjamin, I. *Chemical Physics Letters* **2004**, *393*, 453.

(93) Stephenson, R.; Stuart, J. *Journal of Chemical and Engineering Data* **1986**, *31*, 56.

(94) Sangster, J. *Octanol-water partitioning coefficients: Fundamentals and Physical Chemistry*; John Wiley & Sons: Chichester, U.K., **1997**.

(95) Šegatin, N.; Klofutar, C. *Monatshefte für Chemie / Chemical Monthly* **2004**, *135*, 241.

(96) Haase, R.; Tillmann, W. *Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry & Chemical Physics* **1995**, *192*, 121.

Chapter 3 - The Effect of Urea on Proteins or Peptides using Kirkwood-Buff Theory

3.1 Abstract

It is well known that the stability of a peptide or protein can be influenced by cosolvents in aqueous solution. Here, we use molecular dynamics simulations to investigate how a cosolvent can affect protein/peptide folding and aggregation. We choose urea as the cosolvent and N-methylacetamide (NMA) as a simple peptide model to investigate, with atomic detail, the molecular distributions observed in solution, and how these distributions change with the concentration of urea and NMA. The Kirkwood-Buff (KB) theory of solutions is used to relate changes in the solution thermodynamics to changes in the molecular distributions observed in the solutions, expressed in terms of preferential interactions. The results are then further analyzed in terms of changes in 1st and 2nd solvation shell coordination numbers. It is argued that the simulation of ternary mixtures, coupled with a KB analysis, can provide solid basis for understanding complicated biomolecule and cosolvent interactions.

3.2 Introduction

3.2.1 General Information of Denaturant

Protein denaturation is important for understanding the protein folding and unfolding process. Protein stability can be directly influenced by the addition of cosolvents to an aqueous mixture.[1] Here, cosolvents refer to small molecules other than the primary solvent in a mixed system.[2] Some cosolvents are considered protein denaturants, like urea and guanidinium chloride (GdmCl), they can destabilize protein structure and solubility. Other cosolvents are considered to be osmolytes, like polyols and sugars, and they tend to stabilize the native structure of proteins.[3-8] In addition, it is possible that cosolvents can affect protein interactions and aggregation.[9-16] Hence, it would be useful to understand the molecular interactions between the cosolvent, protein, and water at the molecular level.

From an experimental point of view, the most commonly used means to gather data concerning solvent structure at the atomic level are x-ray diffraction and nuclear magnetic resonance (NMR) experiments.[17] However, both of these approaches have some limitations. X-

ray diffraction can be used to detect ordered water,[18] but not typically ordered cosolvent molecules. NMR studies rarely identify cosolvent molecules.[19-24] Other techniques, like hydrodynamic, small-angle x-ray and neutron scattering, calorimetric, dielectric, and vapor-pressure absorption isotherm experiments, can also provide information at low resolution.[25-28] Unfortunately, many of the cosolvent effects involve multiple weak binding/association effects which do not lead to well defined binding sites for the cosolvent molecules.[1]

Urea is a commonly used protein denaturant. However, the molecular mechanism for the action of urea at atomic level is still not clear. Many studies have investigated the denaturation mechanism from both theoretical and experimental points of view. Some studies have showed that an increase in urea concentration can improve the solubility of both polar and non-polar groups as indicated by the observed transfer free energies.[29,30] Others studies have determined that the urea-amide unit interaction is exothermic, while urea-apolar interaction is generally endothermic.[31-36] A variety of studies aimed at understanding the urea-protein interaction,[37-43] have provided a number of effects that are often difficult to interpret,[44,45] especially at the molecular level.

Up to now, there are two popular ideas concerning the urea denaturation mechanism. One is the direct interaction mechanism, which basically involves urea molecules directly interacting with the protein by specific binding. A larger binding to the denatured state causing the denaturation.[46] The major binding effect involves electrostatic and van der Waals interactions, and binding to the protein backbone is believed to play the major role in denaturation.[29,47,48] In contrast, it is also argued that osmolytes, acting as protein stabilizers, are typically excluded from interacting with the protein backbone. Alternatively, another explanation involves the indirect mechanism. This states that it is the existence of urea that indirectly alters the solvent environment and properties, reducing the hydrophobic effect, and leading to a more favorable denatured state.[49] It emphasizes the effect of urea on water structure, and thereby the exposure of hydrophobic regions leads to destabilization of protein structure.[50-52]

3.2.2 Experimental and Theoretical Study of Urea Denaturation

3.2.2.1 Tanford

In 1962, Charles Tanford developed a model to calculate the difference in free energy between the native and unfolded states of proteins in solution,[53] which suggested the dominant

change comes from the hydrophobic interaction. Meanwhile, he pointed out that the addition of organic substances, like urea and ethanol, would increase the significance of electrostatic interactions as the dielectric constant of solvent is decreased, and that urea has effects on both non-polar and polar side chains of proteins. Further studies on the unfolding of globular proteins in aqueous urea solutions considered the interaction with solvent as the primary reason for protein unfolding.[50] He also provided a quantitative comparison between theoretical and experimental protein unfolding in urea solutions. From his study of several proteins he indicated that the difference in free energy between native and unfolded proteins would be in the range of 100-200 cal/mole per amino acid, and not influenced by the stability of intermediate states.[54,55]

3.2.2.2 Timasheff

Starting from the early 1970s, Serge N Timasheff attempted to discover the mechanism of the cosolvent effect on protein stability and solubility in a three-component system using the concept of preferential interactions.[56-58] He introduced a new physical property called the preferential binding parameter (Γ_{ij} where i and j represent the solute and cosolvent, respectively), and indicated that cosolvent would be denaturants if Γ_{ij} is positive, and implicated the interaction between cosolvent and the protein probably focused on the backbone and aromatic groups. The additives he studied included salts, sugars, and other small organic compounds (urea, glycerol, amino acids). He also indicated that osmolytes would typically increase the surface tension of water, and therefore increase the degree of preferential hydration which will lead to the stabilization of protein structure.[59]

3.2.2.3 Bolen

The studies of Bolen and coworkers have focused on the peptide backbone unit as the most numerous group in proteins.[60,61] Following the original work on glycine, diglycine and triglycine in water and 1 M urea solutions by Nozaki and Tanford,[54] he evaluated transfer free energies for different groups into a variety of cosolvents. A variety of models, such as cyclic glycylglycine, zwitterionic glycine peptides, and N-acetylglycineamide peptides of varying peptide chain length, have been used.[62] His measurements concerning the transfer free energy from water to 1 M urea solutions showed a large transfer free energy for the peptide backbone. Namely, a value of -56 cal/mol for the peptide backbone unit in blocked glycine peptides, and values of -39 and -43 cal/mol for the peptide backbone unit obtained from cyclic glycylglycine

and N-acetylglycinamide models, respectively. This demonstrated a reasonable independency of the transfer free energies on chemical models and suggested a degree of additivity for solvent systems.

3.2.2.4 Murphy

Murphy has studied a homologous series of cyclic dipeptides, e.g. c(AG), c(AA), c(GG) and c(LA), in urea solution of various concentrations at 298 K.[63,64] He determined the thermodynamics of dissolution of cyclic dipeptides into urea solutions. A group contribution to the excess free energy of -73 to -35 J/mol per amide unit peptide in 1 M urea was observed. In addition, he concluded that urea binding to the amide units displaces water in their solvation shell, and therefore urea diminishes the hydrophobic effect which, in turn, leads to protein denaturation. Later, he used also used MD simulation to test the ability of trimethylamine n-oxide (TMAO) to counteract the effect of urea on the stabilization of proteins.[64]

3.2.2.5 Record

Record and coworkers have also studied the thermodynamic properties of the interaction between biomacromolecules and urea. In studies of bovine serum albumin (BSA) they developed a new application of vapor pressure osmometry (VPO)[65] to provide preferential interactions, and demonstrated that urea accumulated around protein surface.[66] In addition, the m-values for protein unfolding have been estimated and compared with the experimental data. The proportionality of the m-value to the protein accessible surface area change (ΔASA) indicated that the difference in preferential binding coefficient ($\Delta\Gamma_{23}$) was proportional to ΔASA.[67,68] In addition, they have studied the urea induced unfolding of the DNA-binding domains of *lac* repressor, and focused on local-bulk partitioning and a competitive binding model to interpret the unfolding mechanism.[69-72] More recently, a detailed comparison between urea and glycine betaine from the viewpoint of preferential interactions illustrated their role as a protein denaturant and stabilizer, respectively.[73,74]

3.2.2.6 Shimizu

Seishi Shimizu has combined Statistical Mechanics and Kirkwood-Buff Theory to help explain the cosolvent denaturation effect. He has estimated the heat capacity (ΔC_p),[75] m-value,[76,77] and excess solvation numbers[78,79] based on preferential interactions and compared

them with experimental data. Using Kirkwood-Buff theory, their theoretical study developed a detailed explanation concerning how urea weakens the hydrophobic effect[80] and explains the preferential interaction.[81]

3.2.3 Simulation Study of Urea Denaturation Mechanism

As an alternative method, people have believed simulation could and will provide more useful information. Here, we briefly review some of the simulations studies of denaturants.

3.2.3.1 Polar, Electrostatic Interactions

Some groups believe that it is the polar and electrostatic interactions between urea and polypeptide chains that have a dominant influence on protein denaturation. O'Brien et al. selected a highly stable helix (H1) from mouse PrP^C (PDB code 1AG2, residues 144-153) as a protein model, and suggested there are strong H-bond interactions between carbonyl oxygen and positive charged solute, even though the overall change on electrostatic interaction is small.[82] Mountain and Thirumalai used a long linear hydrocarbon chain with oppositely charged residues at each end as a model hydrophobic globular protein model. When the two ends of this chain were in contact this was denoted as the folded state, while the breaking of contact was considered as the unfolded state. They observed that the addition of 6 M urea to the solution disrupts the original folded state by a direct interaction between urea and charged chain ends.[83] The same group also presented data suggesting that the valine dipeptide experienced a conformational change in 8 M urea, and concluded the denaturation as a result of a H-bond between urea and the amide backbone.[84] This viewpoint is also supported by Bennion and Daggett, who simulated the urea denatured protein CI2 in solution which indicated H-bond formation with polar residues and the sacrificing of intraprotein interactions.[51]

3.2.3.2 Dispersion Force / Van Der Waals Interactions

Another idea of the denaturation mechanism points to the role of dispersion forces in stabilizing the denatured state. Zangi et al. showed that hydrophobic interactions reduced as urea preferential binding increased when using a hydrophobic chain as a protein model.[85] The contribution to the enthalpy change was related to the strength of the dispersion attraction between urea and the chain, which is more favorable than the attraction between water and the chain. They also believed an indirect mechanism was not sufficient to explain urea

denaturation.[85] Hua et al. simulated hen lysozyme denatured by urea, and showed that water molecules were expelled from the first hydration shell of the protein since the urea-protein interaction was clearly stronger than the water-protein interaction. They indicated that hydrogen bond formation between urea and the protein backbone caused the loss of intra-backbone H-bonds and therefore the loss of secondary structure. Their study supported the direct interaction mechanism as well.[47] Canchi et al. studied the interaction between urea and the Trp-Cage mini-protein using REMD (replica exchange molecular dynamics) simulations, which demonstrated that the denaturation process was mainly driven by van der Waals interactions, while hydrogen bonding was not a dominant role.[86]

3.2.3.3 Hydrophobic Effect

Stumpe et al. suggested that the H-bond between urea-water was much weaker than the H-bond between water-water, and hence the hydrophobic effect drove urea aggregation. The main driving force came from direct interactions between urea and the protein, while indirect effect helped to weaken then hydrophobic effect.[87] In another paper, they computed the energetic change as urea is substituted for water in the solvation shell of protein. They found protein-urea H-bonds were weaker than protein-water or water-water H-bonds, and thus suggested a combined mechanism of hydrophobic and polar interactions.[88] In agreement with previous results, urea was observed to stabilize partially unfolded proteins against hydrophobic collapse. Protein unfolding was not induced by urea active attack, but rather by the more stable equilibrium of thermal fluctuations at unfolded state.[89]

3.2.4 What's next

Obviously, a clear picture of cosolvent effects in biomolecular systems is highly desirable. What is lacking is the detailed accurate information at the atomic level to describe the real mechanism as well as thermodynamic properties. Here we introduce a theory that can provide us information concerning both atomic details and the bulk thermodynamic properties - the Kirkwood-Buff theory of solutions.

Kirkwood–Buff (KB) theory is an exact theory of solutions that relates properties of a solution mixture to radial distribution functions between the different components of the solution.[90] KB theory has been widely used to help understand the basic properties of solutions, the effects of additives on the solubility of solutes and biomolecular equilibria, to investigate the

local composition of solutions in the context of preferential solvation, to interpret computer simulation data, and to develop models for many of the above effects.[77,80,81,91-94]

As the importance and practical applications of KB theory have become apparent, more and more people have shown interest in studying the urea denaturation mechanism by combining KB theory and molecular dynamics simulation.[77,81,92] As part of our effort to develop accurate force fields for the simulation of solution mixtures and their application in biomolecular systems, we recently presented a force field for the simulation of amides.[91] Small amides can be employed as simplified models for peptides and proteins possessing the same functional groups. Here, we choose N-methylacetamide (NMA) to represent a model for the peptide group. In studying peptide and protein denaturation, it is important to maintain a correct balance of hydrogen bonding between the peptide groups, and the degree of solvation of the peptide groups. Too little solvation will tend to favor self-aggregation of the peptide groups, whereas too much solvation will destabilize native state structures. The effect of urea on this balance can be probed using KB theory and simulation. The results can be provided in terms of both molecular detail and thermodynamic properties. This is the approach taken here.

3.3 Theory

The subscripts 1, 2, and 3 refer to the primary solvent (water), the biomolecule solute (NMA), and the cosolvent (urea), respectively. We have embraced the pseudo chemical potential approach of Ben-Naim.[95] This is closely related to the excess chemical potential expressions obtained from statistical mechanics and used in previous simulation studies. However, it is different from the traditional excess chemical potential adopted by experimentalists. In particular, number density (or molarity) is the natural concentration unit for most formulations of the chemical potential used in simulation studies. In contrast, the majority of early experimental studies have focused on the molal concentration scale.

3.3.1 Preferential Binding and Preferential Interactions

Preferential binding is a thermodynamic expression of the degree of cosolvent binding derived for systems open to the solvent and cosolvent. The theory of preferential binding and the concept of preferential interactions[96,97] are aimed to explain cosolvent induced protein denaturation,[55,98] equilibrium dialysis,[97] osmotic stress,[99] the Hofmeister series,[8] and light scattering from protein solutions.[100-102] There are many ways to determine Γ_{23}[2,25,66,97,103-105] and

many applications for cosolvent solutions.[2,25,105] The preferential binding parameter at a temperature T is defined by,[2]

$$\Gamma_{23} = \left(\frac{\partial m_3}{\partial m_2}\right)_{T,\mu_1,\mu_3} \tag{3.1}$$

where m_i is the species molality, T is the temperature, and μ is the chemical potential. The preferential binding is one of the central properties of interest in this work. While it is defined in an open system, we will see that it provides information on cosolvent effects in closed systems of usual interest. It is common to make the following assumption,[97]

$$\left(\frac{\partial m_3}{\partial m_2}\right)_{T,\mu_1,\mu_3} \approx \left(\frac{\partial m_3}{\partial m_2}\right)_{T,P,\mu_3} \tag{3.2}$$

and then according to the thermodynamic relationship

$$\left(\frac{\partial m_3}{\partial m_2}\right)_{T,P,\mu_3} = -\left(\frac{\partial \mu_2}{\partial \mu_3}\right)_{T,P,m_2} \tag{3.3}$$

one can use the above derivatives to quantify preferential binding in closed systems. The addition of a cosolvent changes the chemical potential of the solute which can affect the equilibrium between two forms. Generally speaking, classical denaturants tend to give a large positive preferential binding, while typical osmolytes always show a negative preferential binding which is also referred to as preferential hydration.[92] We will use KB theory to provide a description of the preferential binding in terms of molecular distributions, which can also be obtained from computer simulation.

3.3.2 Kirkwood-Buff Theory

The development of KB theory is described in detail at Chapter 1. The thermodynamic properties of a solution mixture can be expressed in terms of the KB integrals between the different solution components as defined as[95]

$$G_{ij} = 4\pi \int_0^\infty \left[g_{ij}^{\mu VT}(r) - 1\right] r^2 dr \tag{3.4}$$

where, G_{ij} is the KB integral between species i and j in the solution mixture, g_{ij} is the corresponding radial distribution function (rdf) in the μVT ensemble, and r is the center of mass to center of mass distance. KB integrals were determined from the present simulation data (NpT ensemble) by assuming that,[90,94]

$$G_{ij}(R) = 4\pi \int_0^R \left[g_{ij}^{NpT}(r) - 1\right] r^2 dr \tag{3.5}$$

where R represents a correlation region within which the solution composition differs from the bulk composition. All rdfs are assumed to be unity beyond R. Excess coordination numbers are defined as

$$N_{ij} = \rho_j G_{ij} \tag{3.6}$$

where a value of N_{ij} greater than zero indicates an excess of species j in the vicinity of species i (over a random distribution), while a negative value corresponds to a depletion of species j surrounding i. The KB integrals G_{ij} typically display large variations with composition, and minor modification of the force field parameters also lead to large deviations. Hence, this is a very sensitive test of the quality of a force field.

No approximations are made during the derivation of the KB related equations. Previous studies have indicated that a combination of KB theory and NpT simulations can provide quantitative information concerning the thermodynamics of solutions[94] The KB integrals, together with the corresponding excess coordination numbers, have provided a simple physical picture of changes in the local solution composition around each species.

We believe that the KB integrals, which can be obtained from experiment or through simulation, provide the most promising approach to improve our understanding of cosolvent effects. Therefore, one obtains the derivatives for a ternary system

$$m_i = \frac{\rho_i}{\rho_1} = \frac{n_i}{n_1} \tag{3.7}$$

$$\rho_j = \frac{n_j}{V} \tag{3.8}$$

where m_i is the species molality, n_i is the number of molecules of i, and ρ_i is the number density.

3.3.3 Application of KB Theory to Three Component Mixtures

Expressions for the preferential interactions, the chemical potential derivatives and associated activity derivatives in terms of KB integrals are then given by the following expressions[92,106,107]

$$N_{ij}^{+} = N_{ij} + m_j\left(1 + N_{11} - N_{i1} - N_{j1}\right) \tag{3.9}$$

$$\Gamma_{ij} = \rho_j\left(G_{ij} - G_{i1}\right) \tag{3.10}$$

$$\mu_{ij} = \beta\left(\frac{\partial \mu_i}{\partial ln m_j}\right)_{T,P,m_{i\neq j}} \tag{3.11}$$

$$\mu_{33} = \beta\left(\frac{\partial \mu_3}{\partial ln m_3}\right)_{T,P,m_2} = \frac{(1 + N_{22}^{+})}{(1 + N_{22}^{+})(1 + N_{33}^{+}) - N_{23}^{+}N_{32}^{+}} \tag{3.12}$$

$$\mu_{23} = \beta\left(\frac{\partial \mu_2}{\partial ln m_3}\right)_{T,P,m_2} = -\frac{N_{23}^{+}}{(1 + N_{22}^{+})(1 + N_{33}^{+}) - N_{23}^{+}N_{32}^{+}} \tag{3.13}$$

$$\Delta\Delta G_{tr} = \beta\left(\frac{\partial \mu_2^{*}}{\partial m_3}\right)_{T,P,m_2} = -\frac{\Gamma_{22}}{m_3}\mu_{23} - \frac{\Gamma_{23}}{m_3}\mu_{33} \tag{3.14}$$

where β = 1/RT. The corresponding partial molar volume expressions reduce to

$$\rho_2\bar{V}_2 = m_2\frac{(1 + N_{11} - N_{21})(1 + N_{33}^{+}) - (1 + N_{11} - N_{31})N_{23}^{+}}{(1 + N_{22}^{+})(1 + N_{33}^{+}) - N_{23}^{+}N_{32}^{+}} \tag{3.15}$$

$$\rho_3\bar{V}_3 = m_3\frac{(1 + N_{11} - N_{31})(1 + N_{22}^{+}) - (1 + N_{11} - N_{21})N_{32}^{+}}{(1 + N_{22}^{+})(1 + N_{33}^{+}) - N_{23}^{+}N_{32}^{+}} \tag{3.16}$$

$$\rho_1\bar{V}_1 = 1 - \rho_2\bar{V}_2 - \rho_3\bar{V}_3 \tag{3.17}$$

and the volume fraction is given by,

$$\phi_i = \rho_i V_i = \frac{n_i V_i}{\sum n_i V_i} \tag{3.18}$$

3.4 Methods

3.4.1 KBFF Models

All molecular dynamics simulations were performed using our newly developed Kirkwood-Buff force fields for the NMA[91] and urea[93] model developed by Smith group together with the SPC/E water model.[108] Three components mixtures were studied with different concentrations of NMA at 0, 1, 2, 4 molality, and urea at 0, 2, 4, 6, 8, 10 molality respectively; a total of 24 different systems. Every system was simulated in a 6 nm cubic box.

3.4.2 Molecular Dynamic Simulations

All simulations were performed in the isothermal isobaric ensemble at 300K and 1 atm using GROMACS package.[109] The Berendsen weak coupling technique was used to modulate the temperature and pressure with relaxation times of 0.1 and 0.5 ps, respectively, and 4.5×10^{-5} bar^{-1} as compressibility.[110] A time-step of 2 fs was used and the bond lengths were constrained using Settle and LINCS algorithms for water and non-water molecules.[111,112] The particle-mesh-Ewald technique was used to evaluate electrostatic interactions.[113] A real space convergence parameter of 3.5 nm^{-1} was used in combination with twin range cutoffs of 1.0 and 1.5 nm, and a non-bonded update frequency of 10 steps. The steepest descent method was then used to perform energy minimization. This was followed by extensive equilibration, which was continued until all intermolecular potential energy contributions and rdfs displayed no drift with time. Configurations were saved every 1.0 ps for analysis. All mixtures were run for 50 ns simulations; the final 10 ns were used for calculating ensemble averages.

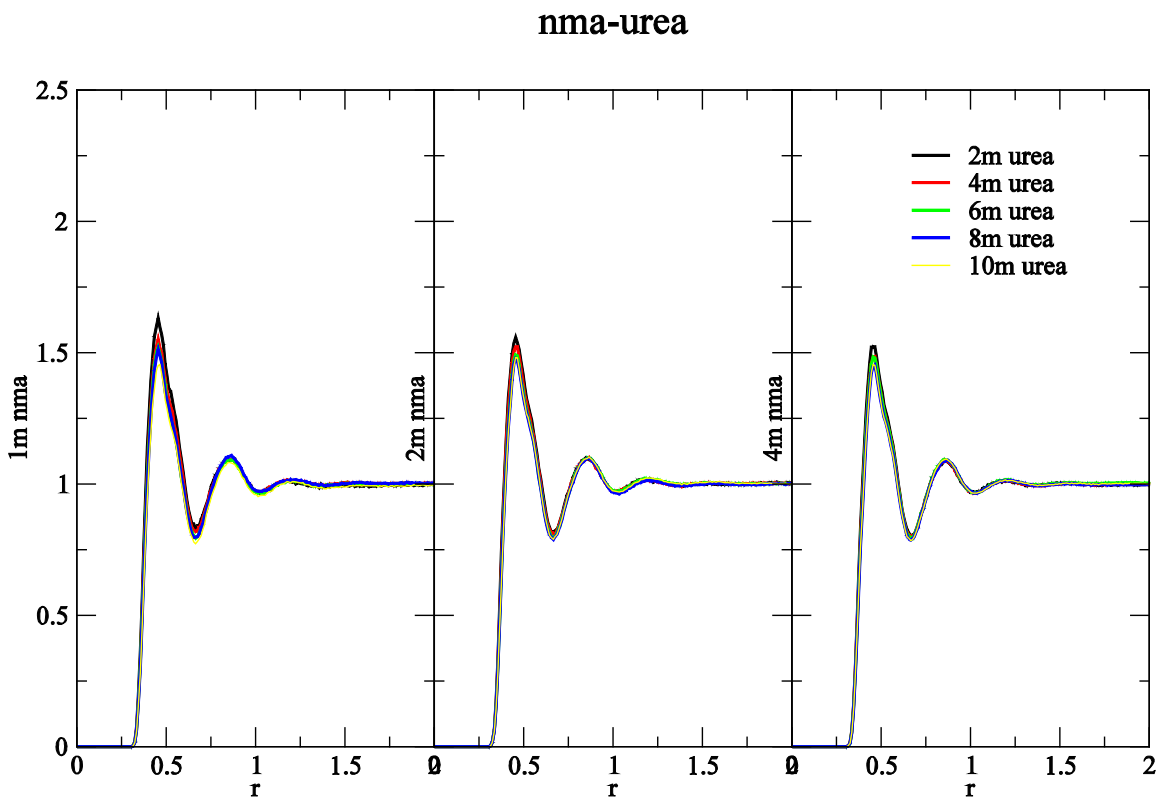
3.5 Result and Discussion

3.5.1 KB Analysis

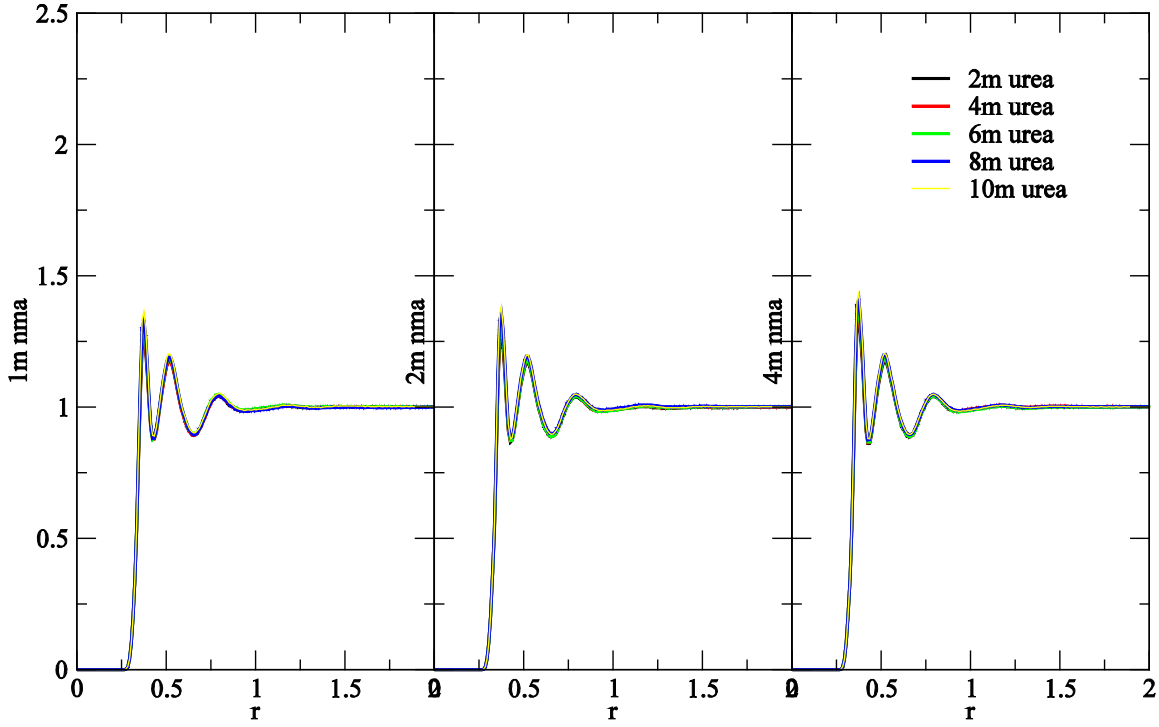
Center of mass based radial distribution functions (rdf) obtained for mixtures with a series of NMA and urea concentration are displayed in Figure 3.1. The rdfs indicated that the first shell solute-cosolvent (NMA-urea) interaction decreased as NMA or urea concentration

increased; while the NMA-H_2O and urea-H_2O interactions were nearly identical as the molality changed. All rdfs showed slight fluctuations in value at distance above 1.5nm. Hence, the KB integrals for all systems studied here were obtained from the simulated data by averaging the values from 1.5 nm to 2.0 nm. Positive numbers generally indicate a net attraction between two species, while negative numbers indicate the two species did not display a favorable interaction.

Figure 3.1 Simulated center-of-mass radial distribution functions (g_{ij}) as a function of distance (r).



nma-h2o



urea-h2o

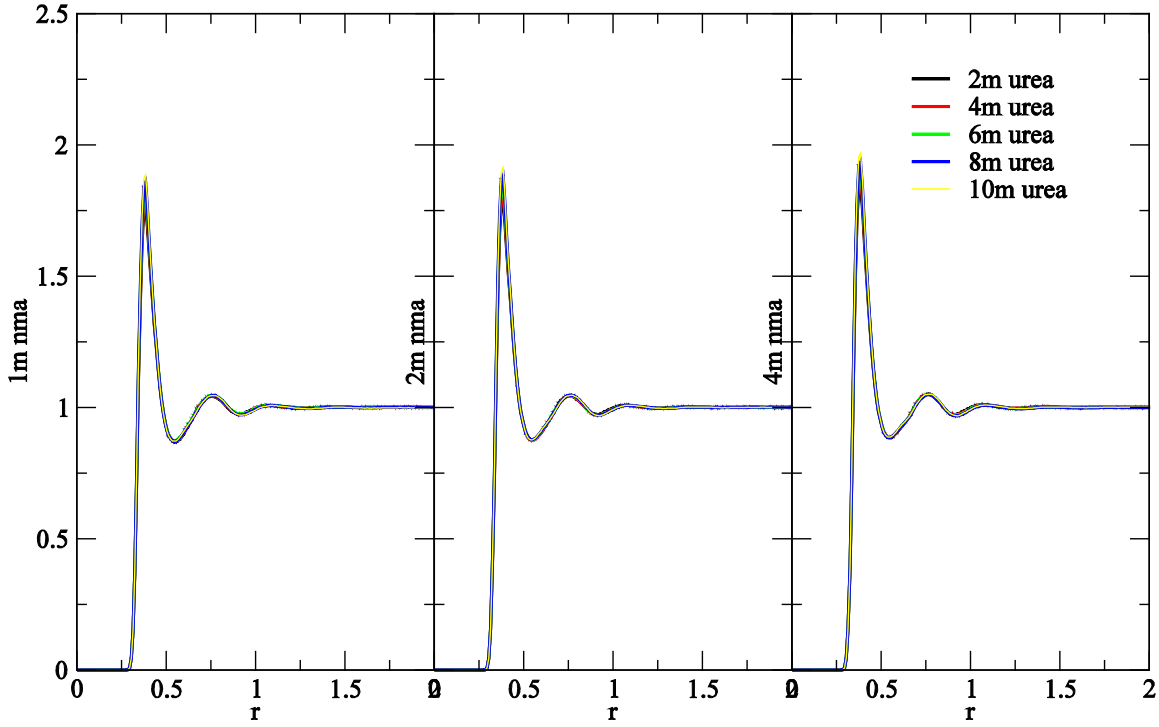
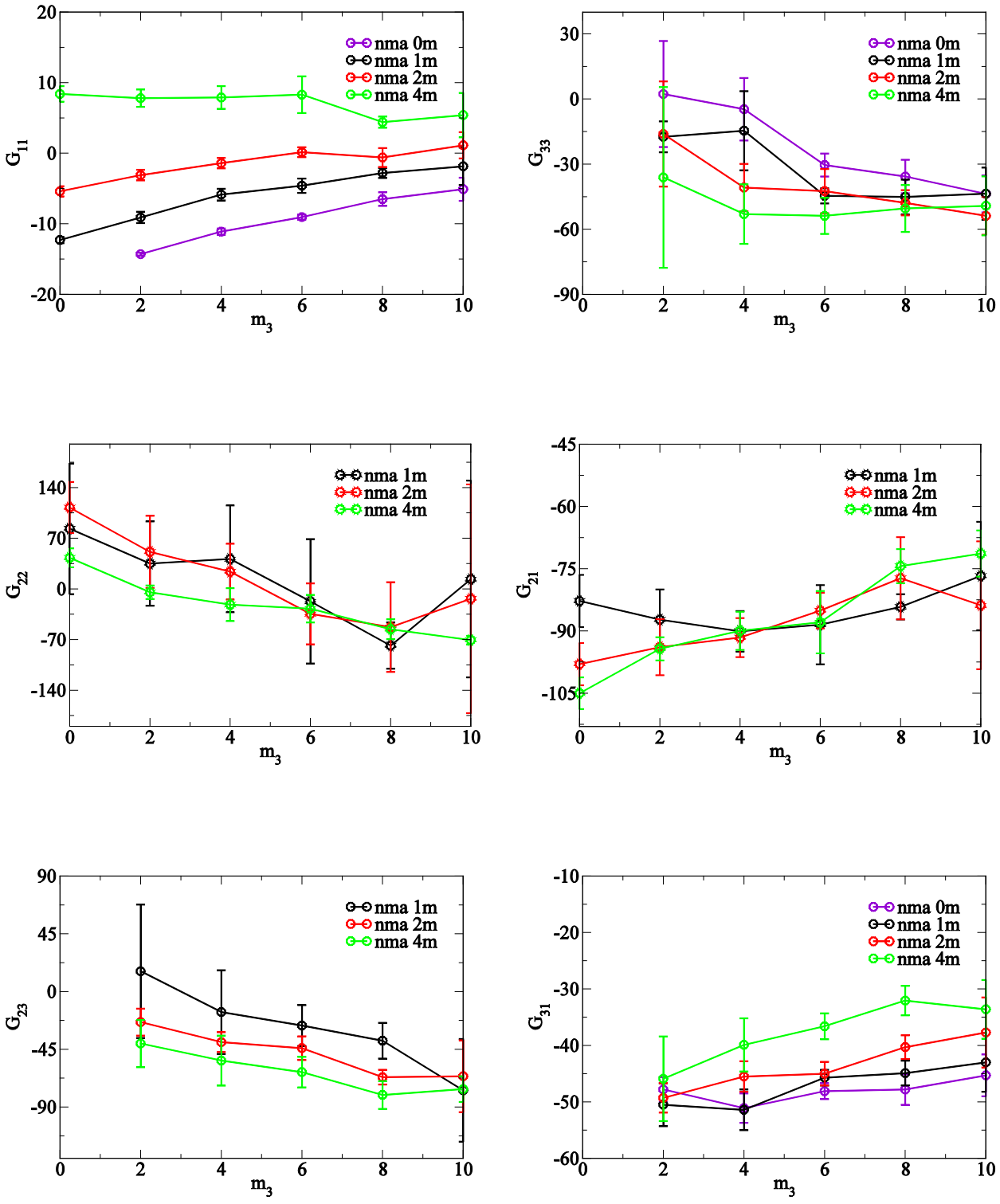


Figure 3.2 Simulated KB integrals (G_{ij}) as a function of urea molality.

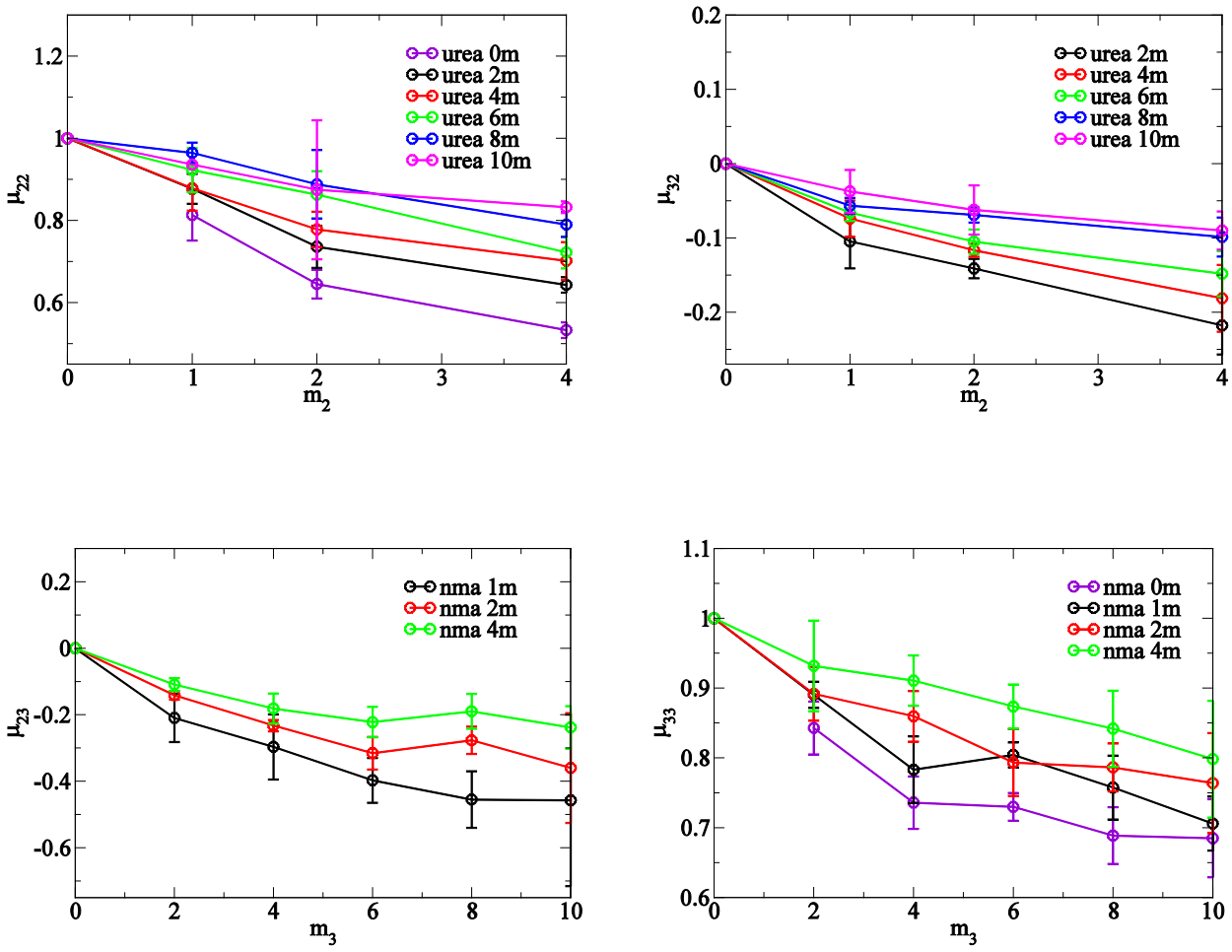


From Figure 3.2 one observes that as the NMA and urea concentrations increased, both G_{11} and G_{31} increased, which indicates that the number of water molecules surrounding other water molecules, or other urea molecules, increases – presumably due to an increase in favorable interactions between these species. Moreover, G_{11} changed from negative to positive values. On the other hand, G_{22}, G_{23} and G_{33} decreased from positive to negative respectively, suggesting NMA and urea lost the preference to cluster together, with either with its own species or with each other. G_{21} slightly increased as the urea concentration increased, but did not display an obvious trend with NMA concentration. Besides, it remained in the range from -110 to -70, which indicates that NMA and water are repulsive to each other, although the addition of urea could alleviate the unfavorable interaction to some degree.

3.5.2 Chemical Potentials and Preferential Binding

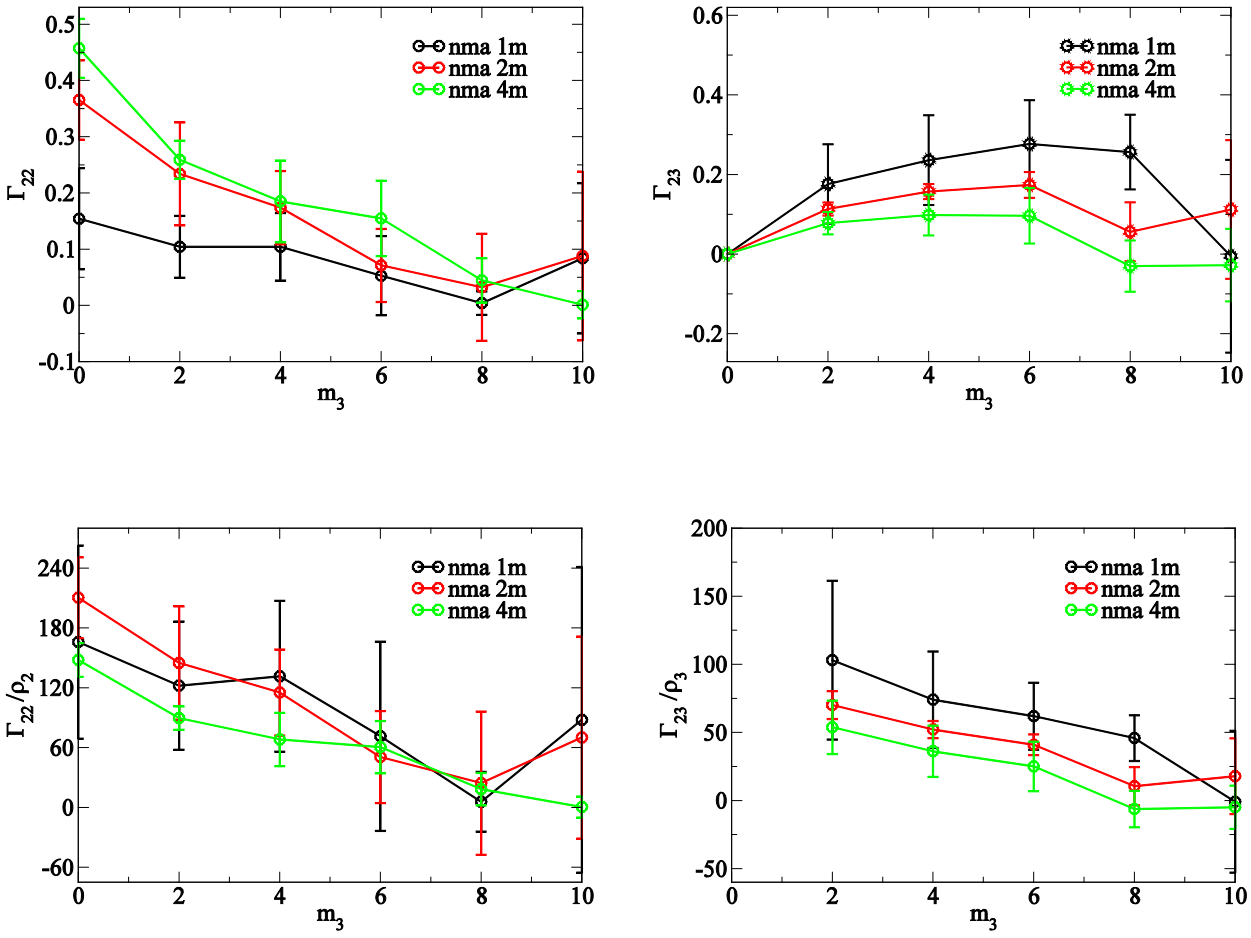
From Equation 3.11, 3.12 and 3.13, we can calculate the chemical potential changes in three components systems by a combination of the above KB integrals. Figure 3.3 displays the results for μ_{22} and μ_{32} as a function of NMA molality, together with μ_{23} and μ_{33} as a function of urea molality. We observe that as NMA increases and urea decreases, the values of μ_{22} and μ_{32} decrease, but the values of μ_{23} and μ_{33} increase. This followed the trend that μ_{ij} decreases as the species j's concentration increases. On the other hand, μ_{22} and μ_{33} are positive numbers while μ_{23} and μ_{32} are negative numbers – as they should be for stable solutions. The effect of increasing urea concentration was to decrease the chemical potential of the NMA solute, i.e. this is a stabilizing situation. Interestingly, the effect is larger for smaller NMA concentrations. We will see that the value of μ_{33} is of central importance in our understanding of cosolvent binding to, and exclusion from, biomolecules in solution.

Figure 3.3 Simulated chemical potential values (μ_{ij}) as a function of nma and urea molality.



According to Equation 3.10 the preferential binding (Γ) can be expressed in terms of combinations of KBIs. The preferential binding was calculated and is presented in Figure 3.4 as a function of urea molality. As mentioned previously, if $\Gamma_{23} > 0$ then the cosolvent displays denaturant properties, while if $\Gamma_{23} < 0$ the cosolvent resembles the properties of an osmolytes. It is not difficult to tell from this figure that $\Gamma_{23} > 0$, which indicates that the cosolvent in this study (urea) is acting as a denaturant. This is to be expected, but is an important observation if we are to use NMA as a model for the peptide group. The preferential interaction decreases with increasing urea concentration, and is larger for smaller NMA concentrations.

Figure 3.4 The simulated preferential binding parameter (Γ_{ij}) for urea and nma at 300K as a function of urea molality.



3.5.3 Contribution to the Preferential Binding

Equation 3.10 indicates how to relate the KB integrals to preferential binding. It is interesting to decompose the Γ values to investigate the contribution from each shell of the radial distribution function. An example of this is provided for NMA at 4 m and urea at 2 m as shown in Figure 3.5. The black line indicates the rdf between NMA and urea, which was used to

determine a series of solvation shells for urea around NMA, while the red rdf line indicates the corresponding NMA-H_2O rdf. According to the shape of the NMA and urea rdf we decomposed Γ_{23} into four parts indicated as the excluded region (blue), first solvation shell (green), and second solvation shell (yellow).

Figure 3.5 The decomposition of Γ_{23} and g_{23} as a function of distance r at NMA 4 molality and urea 2 molality.

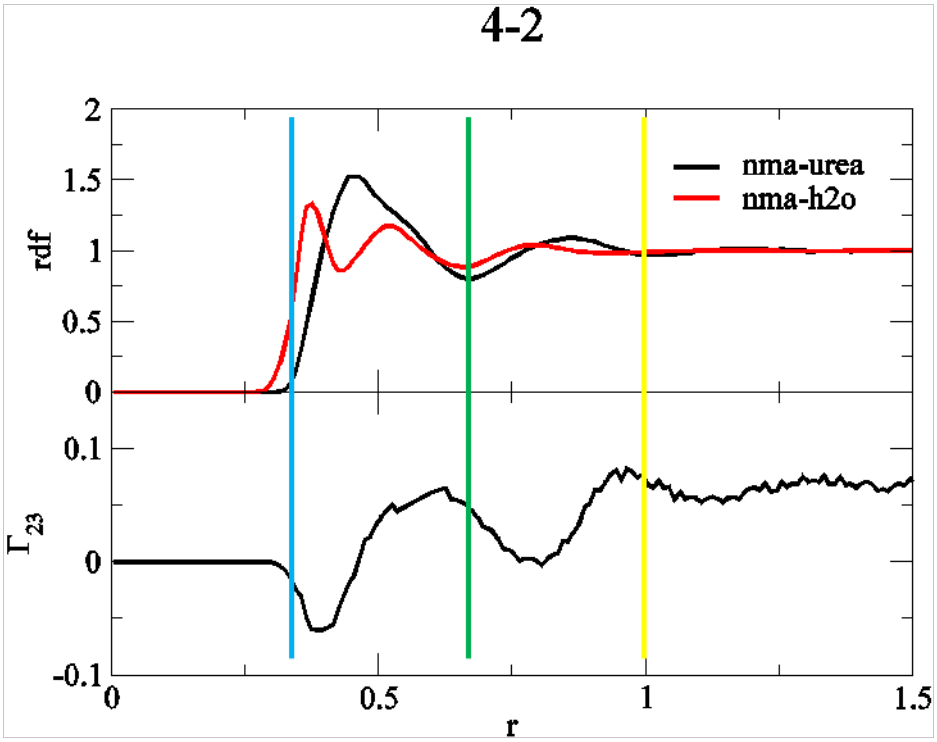
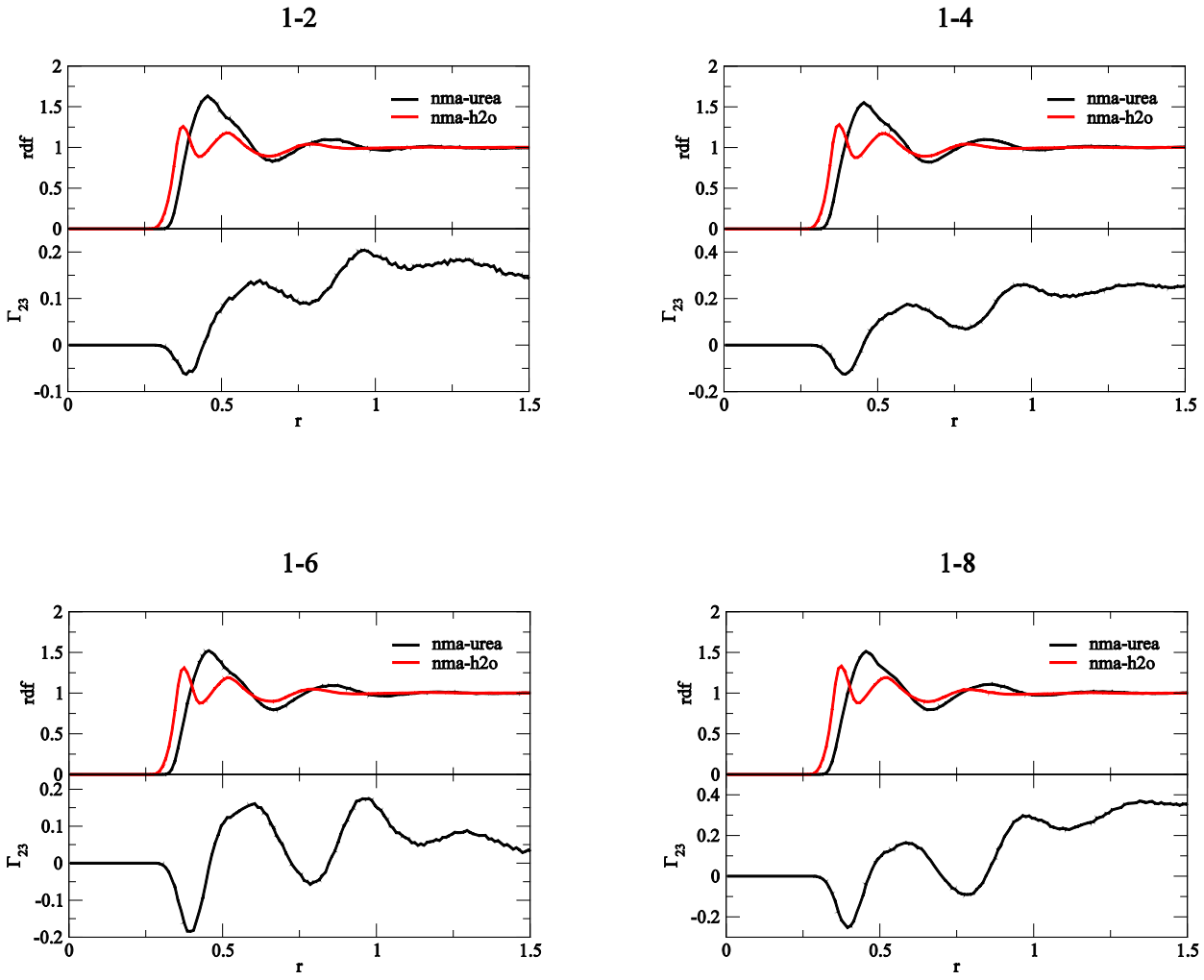


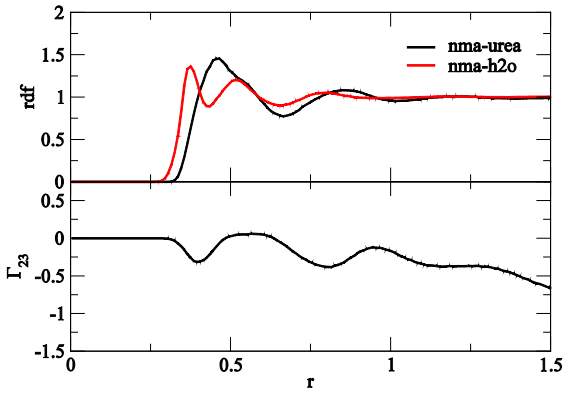
Figure 3.6 provides the comparisons of the rdf and Γ_{23} values for all the compositions of NMA and urea studied here. From these figures, we can observe the following common characteristics of Γ_{23}. The values start from zero and then become negative since g_{21} was non-zero and g_{23} is initially zero, i.e. there is an excluded region where the smaller water molecules can contact the solute, but the larger urea molecules cannot. As g_{23} then increases, Γ_{23} increases

and changes from negative to positive number, generally within the first shell of Γ_{23}. As g_{23} went through the second shell and further, Γ_{23} experienced additional peaks and kept fluctuating.

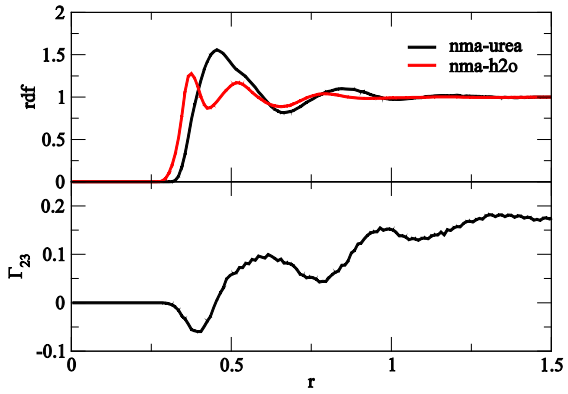
Figure 3.6 Simulated center of mass RDFs g_{23} (black) and g_{21} (red) at different nma-urea molality composition corresponding to respective Γ_{23} as a function of distance r.



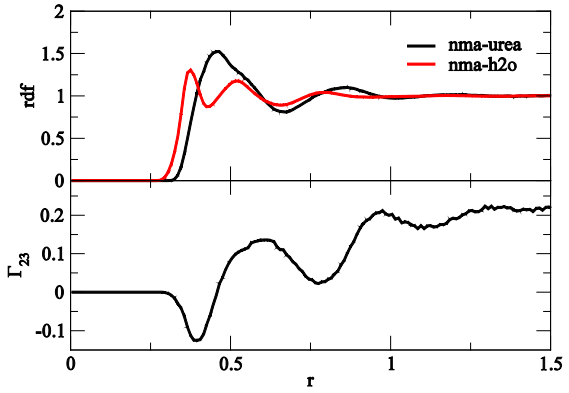
1-10



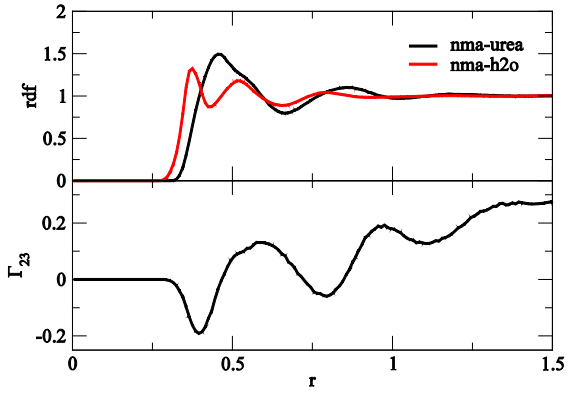
2-2



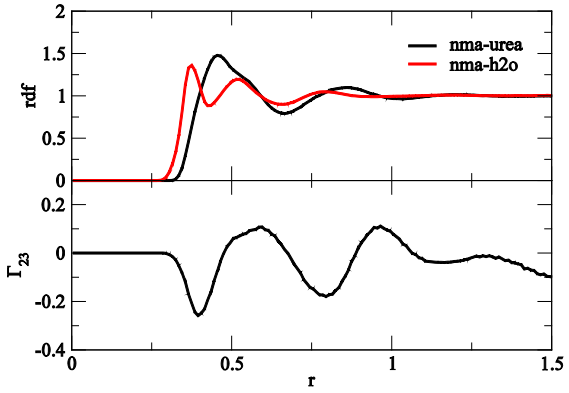
2-4



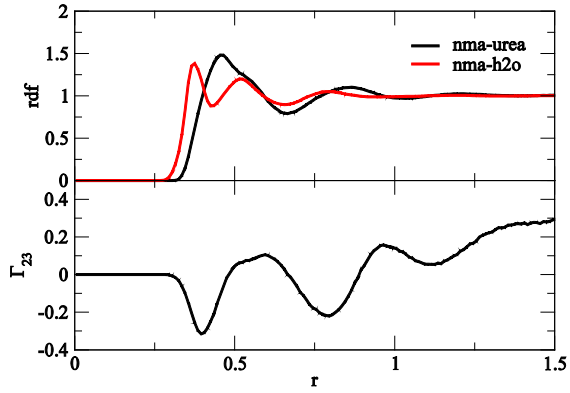
2-6



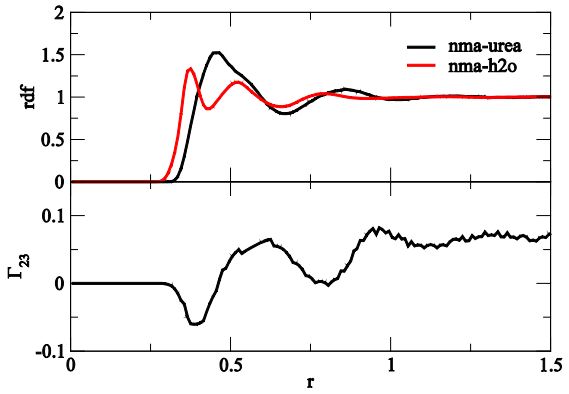
2-8



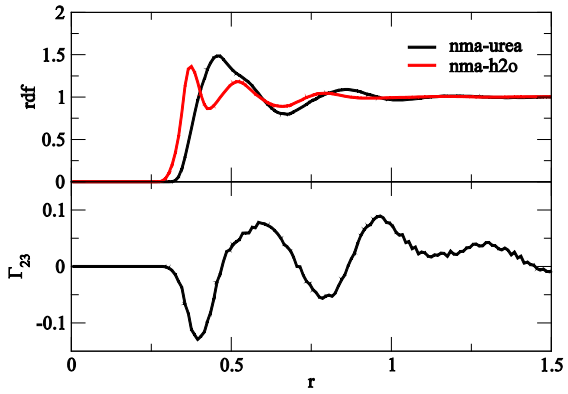
2-10



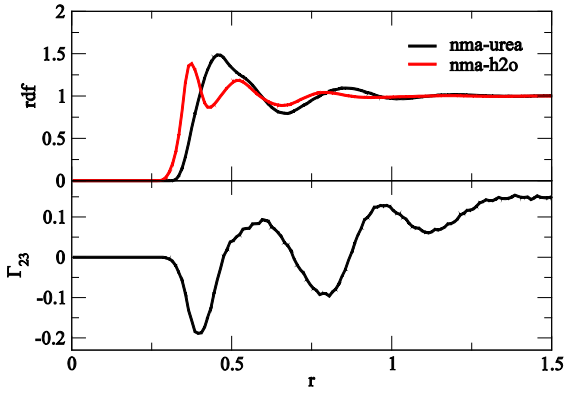
4-2



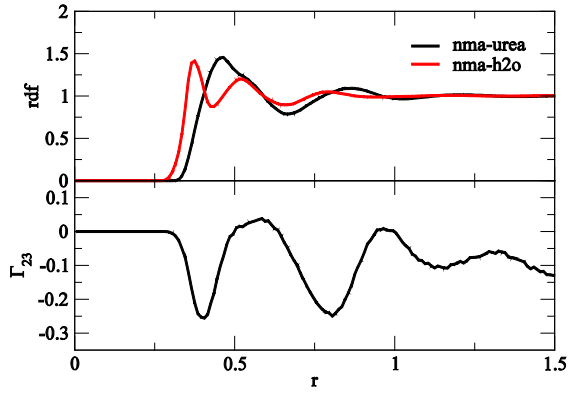
4-4



4-6



4-8



4-10

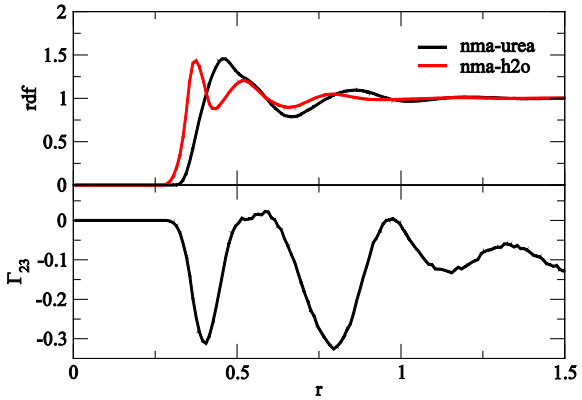
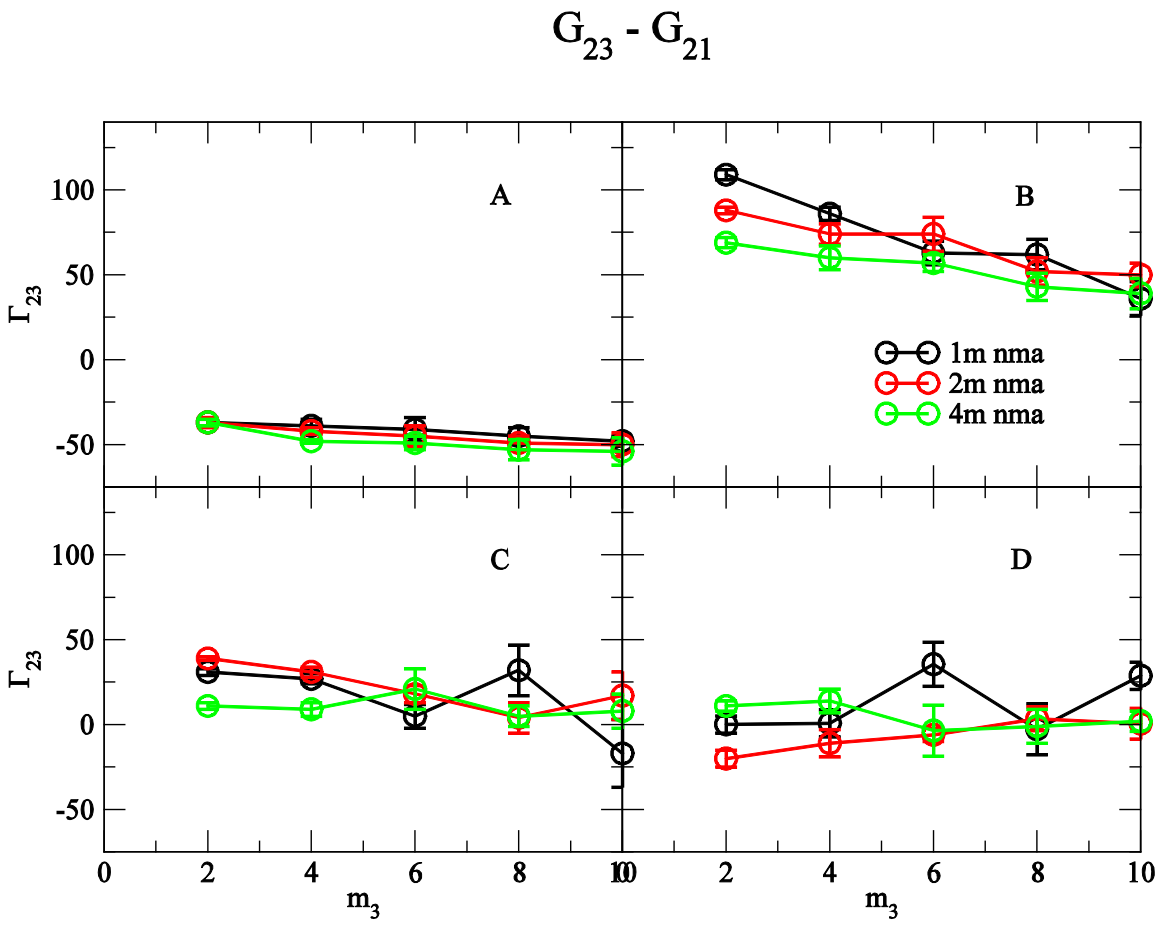


Figure 3.7 The separation of (G_{23} - G_{21}) cm^3/mol into four distance dependent terms as a function of urea molality according to the relative radial distribution functions. A) r = 0 ~ 0.3 nm where g_{23} stays at zero. B) r = 0.3 ~ 0.67 nm where g_{23} is in 1st shell. C) r = 0.67 ~ 1.02 nm where g_{23} is in 2nd shell. D) r = 1.02 ~ 1.5 nm or residual g_{23}.



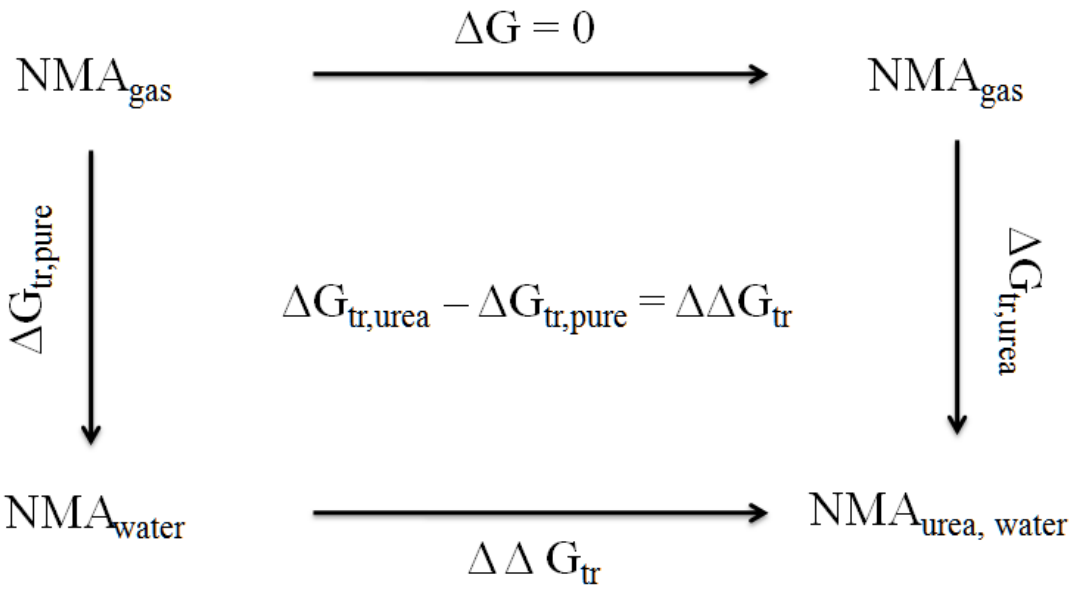
The cutoffs for each region of the preferential interactions were calculated from the average distance before and after each shell. These were determined to be 0.3 nm, 0.67 nm and 1.02 nm. In Figure 3.7 we display the value of each term as a function of urea concentration. It also showed that this didn't influenced too much by NMA concentration change. From 0 to 0.3 nm, G_{23} - G_{21} was a negative number around -50. From 0.3 to 0.67 nm, G_{23} - G_{21} was a large

positive number 50 ~ 100. From 0.67 to 1.02 nm, G_{23} - G_{21} was a small positive number 0 ~ 50. From 1.02 to 1.5 nm, G_{23} - G_{21} is leftover around 0. Overall, the second portion, which corresponds to the first urea shell of g_{23} made the largest contribution to the whole preferential binding.

3.5.4 Transfer Free Energies of NMA to Urea Solutions

The thermodynamics of transfer for the NMA solute to various urea solutions can be determined from the simulations using the thermodynamic cycle provided in Figure 3.8. The figure indicates the transfer free energy of NMA in to urea ($\Delta\Delta G_{tr}$), which can be calculated by Equation 3.14. A negative number indicates a favorable process.

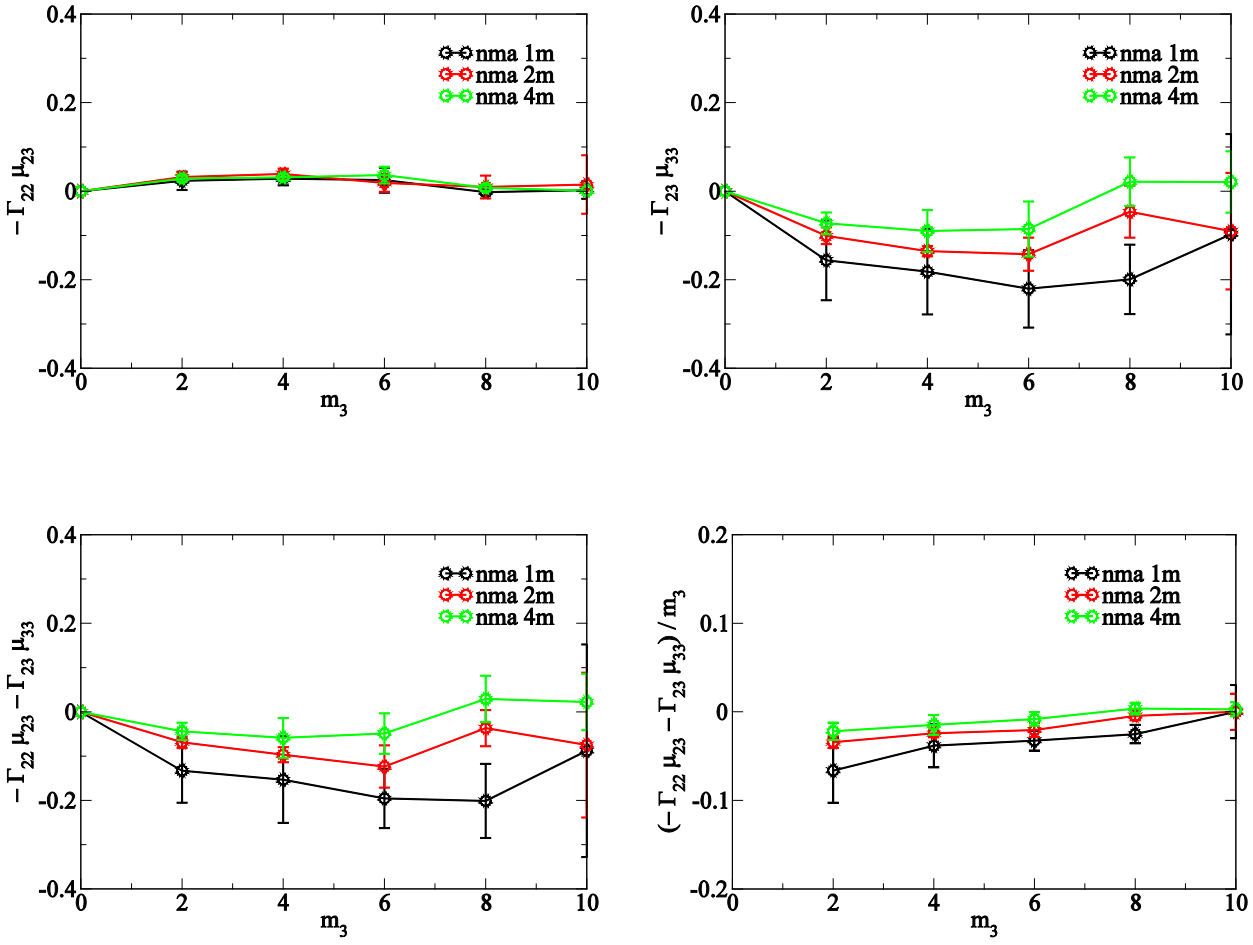
Figure 3.8 The transfer free energy cycle.



According to Equation 3.14, and the previous results concerning the chemical potential and preferential binding, we know that $\mu_{23} < 0$, $\mu_{33} > 0$, $\Gamma_{22} > 0$, $\Gamma_{23} > 0$. Therefore $-\Gamma_{22}\mu_{23} > 0$ and $-\Gamma_{23}\mu_{33} < 0$. The sum of these two terms divided by the urea molality is the change in transfer free energy with urea concentration. Thus, the nature of the transfer free energy change

(favorable or not) is determined by which one of these two terms is dominant. Figure 3.9 provides a detailed thermodynamic analysis of this process.

Figure 3.9 The simulated transfer free energy of NMA in urea solutions and a comparison of the two terms contributing to the transfer free energy.



From above figure it is observed that the first term $-\Gamma_{22}\,\mu_{23}$ was only slightly greater than 0, while the second term $-\Gamma_{23}\,\mu_{33}$ was a relatively large negative number compared the first term. Hence, the second term was dominant for the range of transfer free energies. Furthermore, a linear trend with urea concentration is observed for the transfer free energy. This allows us to

estimate the transfer free energy $\Delta\Delta G$ of NMA to 1 m urea at 298K as -90 ~ -30 J/mol depending on the NMA concentration. When compared to the experimental result of Murphy,[63] which was -73 ~ -35 J/mol using amide unit, and Bolen,[62] which was -300 ~ -60 J/mol using peptide backbone unit, our result appears very reasonable.

3.5.5 Other Thermodynamic Properties

KB integrals and computer simulation data can be applied to connect with additional thermodynamics properties as well.

Figure 3.10 The simulated enthalpy of mixing as a function of urea molality.

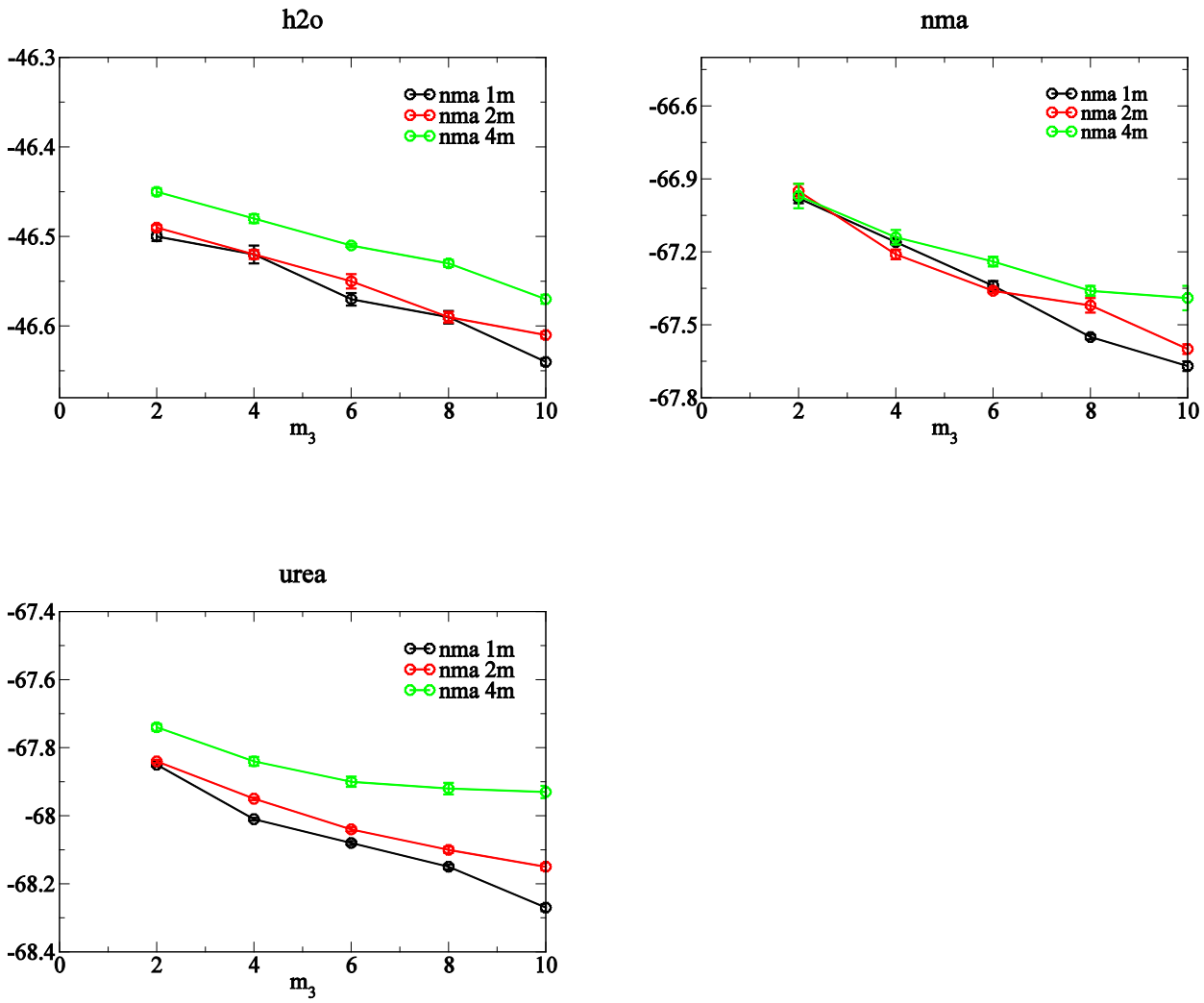


Figure 3.10 provides the enthalpy of mixing of water, NMA and urea respectively. The enthalpy of mixing of water, NMA and urea all decreased as the NMA concentration decreased and the urea concentration increased. Even though there are not experimental data on the full ternary system available to compare with our result, we can extrapolate and estimate the enthalpy of mixing value at 0 m NMA and/or 0 m urea, which are -46.4 kJ/mol close to experimental results -44.0 kJ/mol.[108]

Figure 3.11 The simulated partial molar volume as function of urea molality.

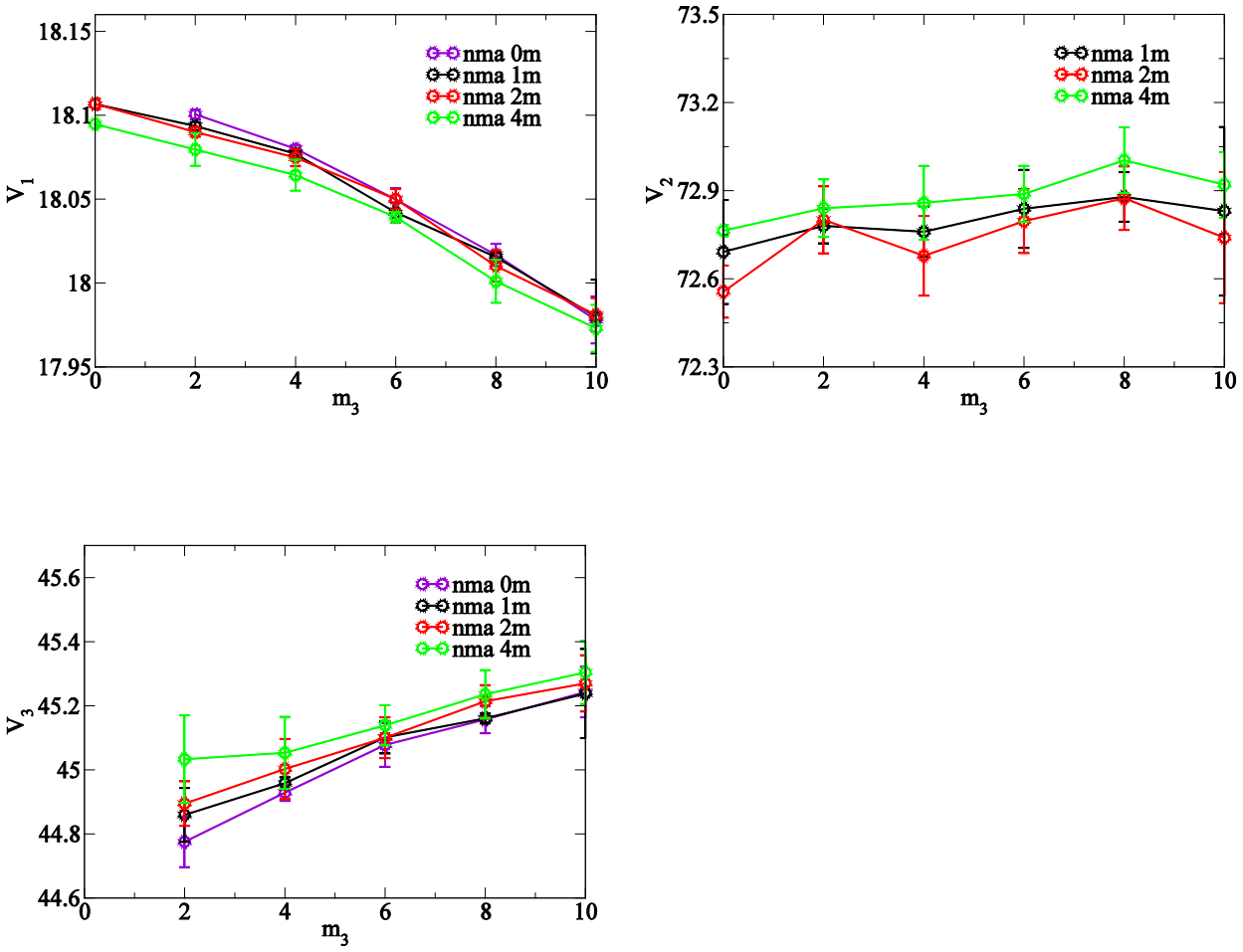


Figure 3.11 displays the simulated partial molar volumes as a function of urea molality. As the NMA and urea concentration increased, the partial molar volume of both NMA and urea

increased, but the partial molar volume of water decreased. Again, at when the molality of urea and nma is zero, we obtain the simulated partial molar volume of pure water is about 18.3 cm^3/mol, which is very close to experimental data 18.0 cm^3/mol.[114]

Figure 3.12 The simulated volume fraction as function of urea molality.

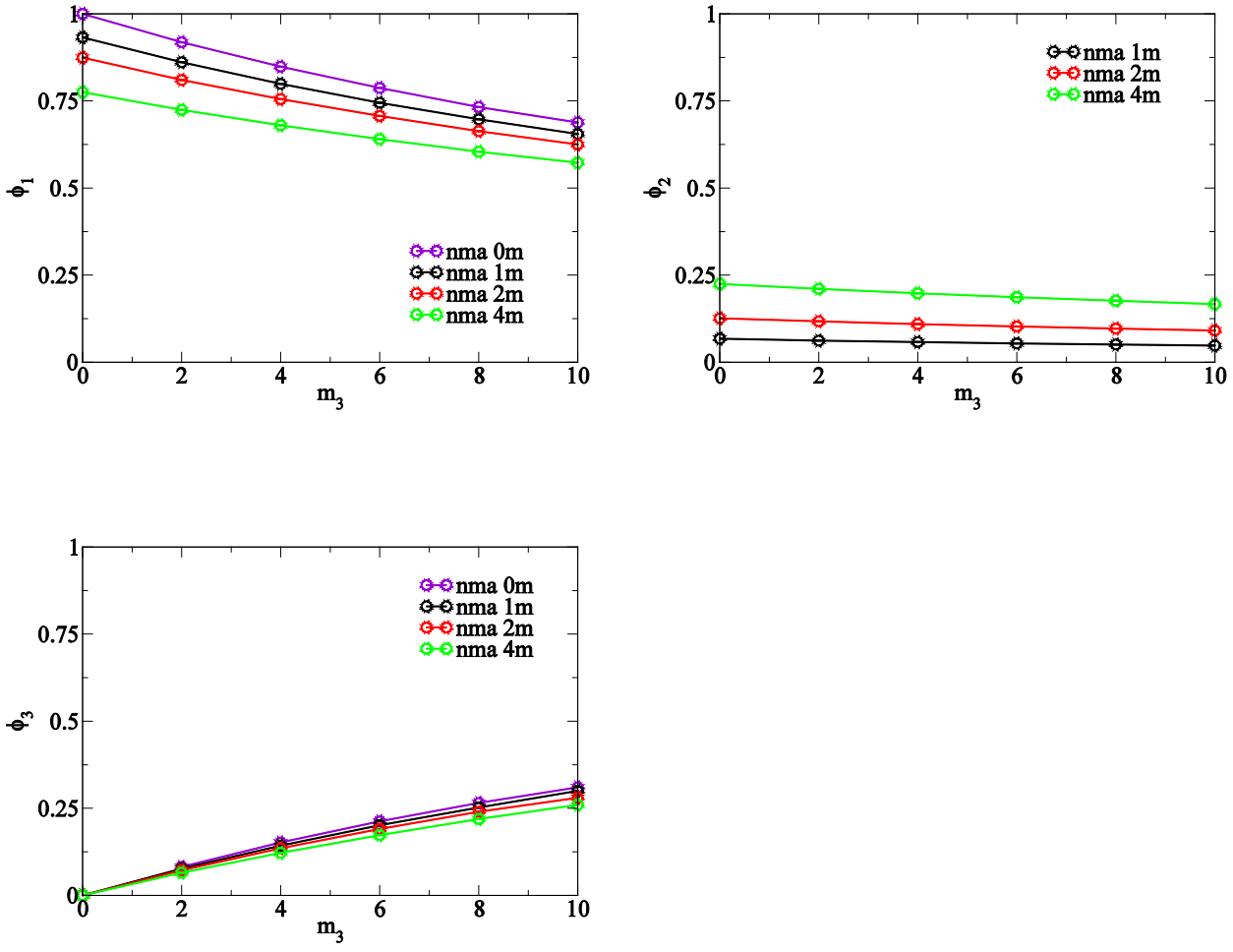


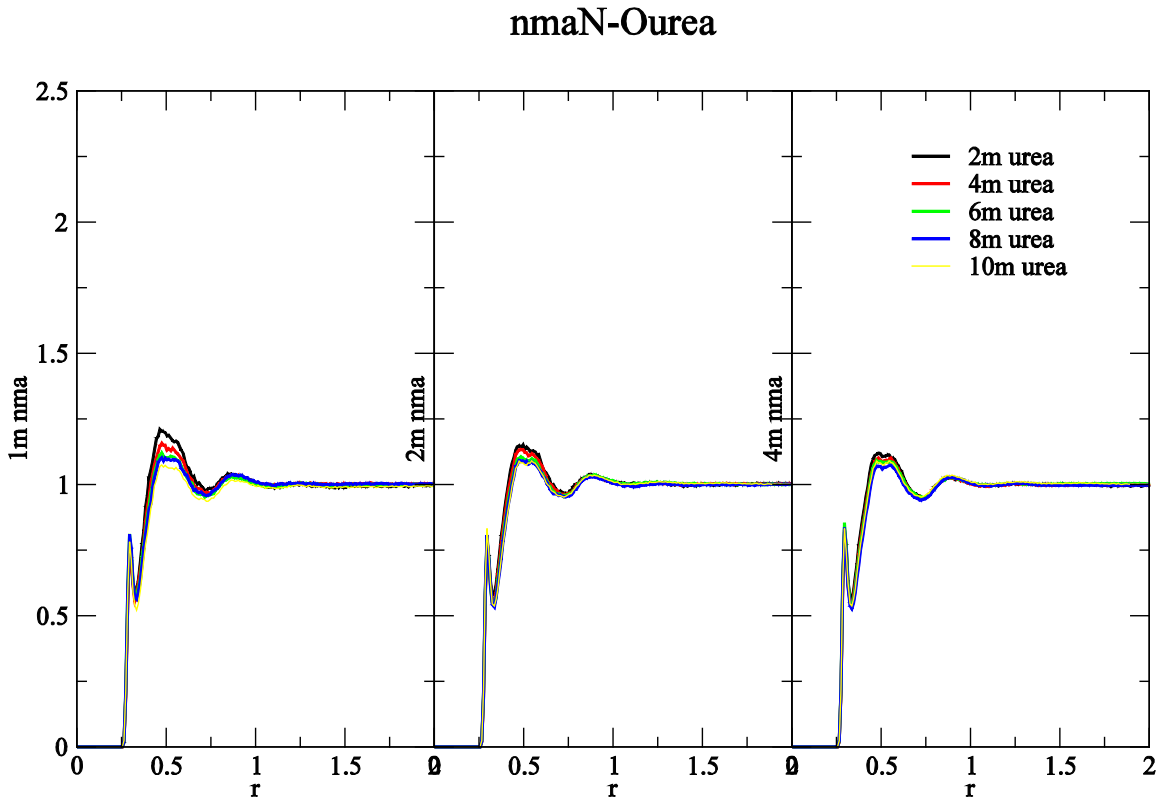
Figure 3.12 shows the simulated volume fractions as a function of urea molality. The volume fraction of NMA increased as the NMA concentration went up, and decreased a little as urea concentration increased. Similarly, the volume fraction of urea increased as the urea

concentration increased, and slightly decreased as NMA concentration increased. The volume fraction of water went down as NMA and urea concentration increased.

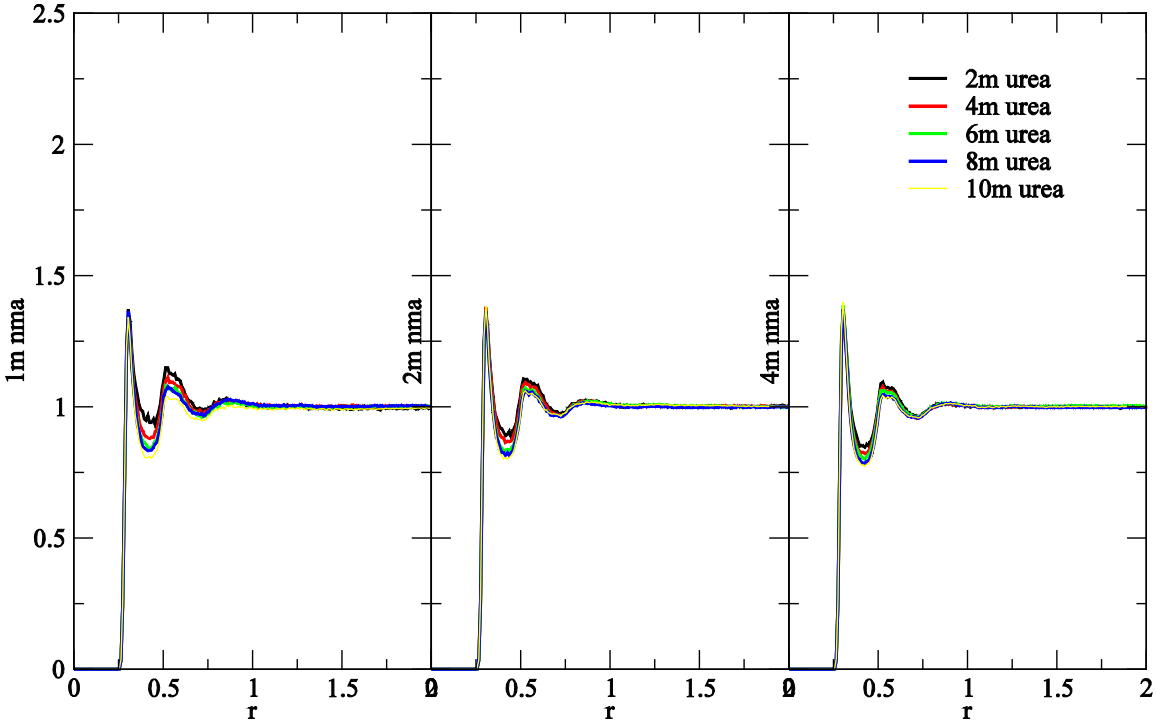
3.5.6 Atom-Atom Radial Distribution Function

Atom-atom based radial distribution functions can probably provide us more details concerning the formation of H-bond in cosolvent solution at different concentrations. Possible H-donors are N-H from NMA, O-H from water, N-H from urea, while possible H-acceptors are O from water O from NMA, O from urea. In Figure 3.13, we display the atom-atom radial distribution functions that correspond to hydrogen bonding, with the first peak indicating a significant H-bond interaction. We observe a slight change with urea and NMA concentration, but any trends among the different molality concentrations of urea and NMA are difficult to determine.

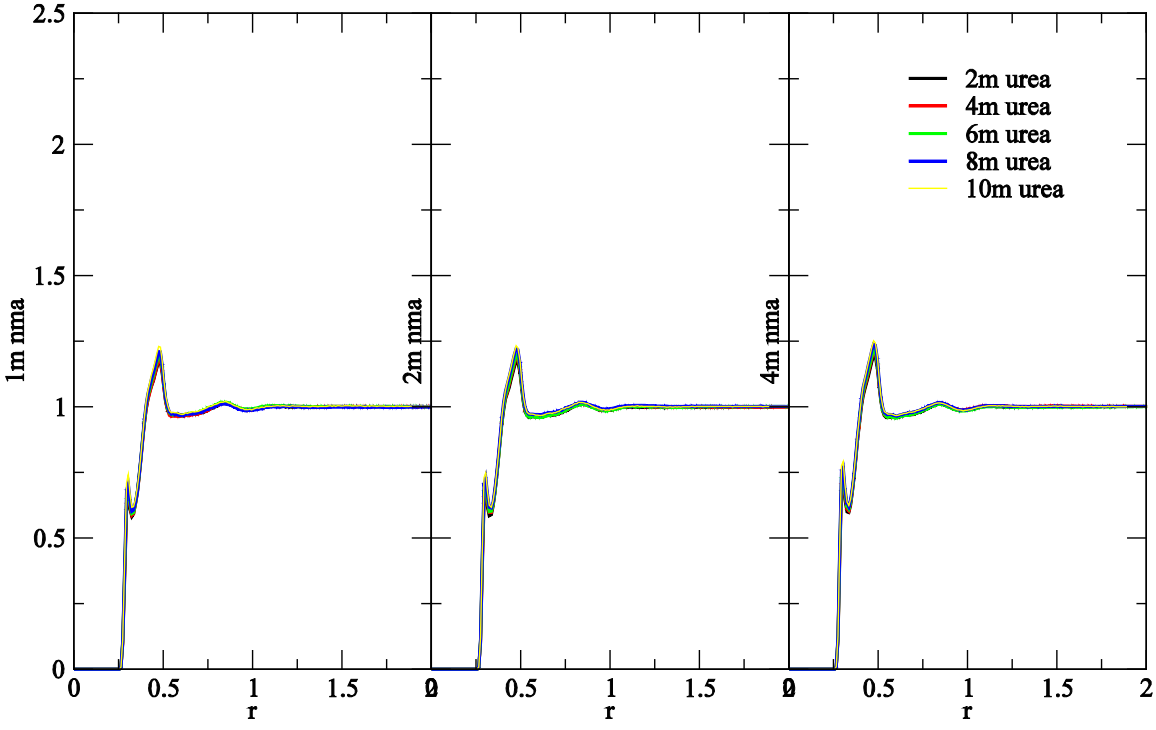
Figure 3.13 The simulated atom-atom radial distribution functions (g_{ij}) as a function of distance *r* at 300 K.



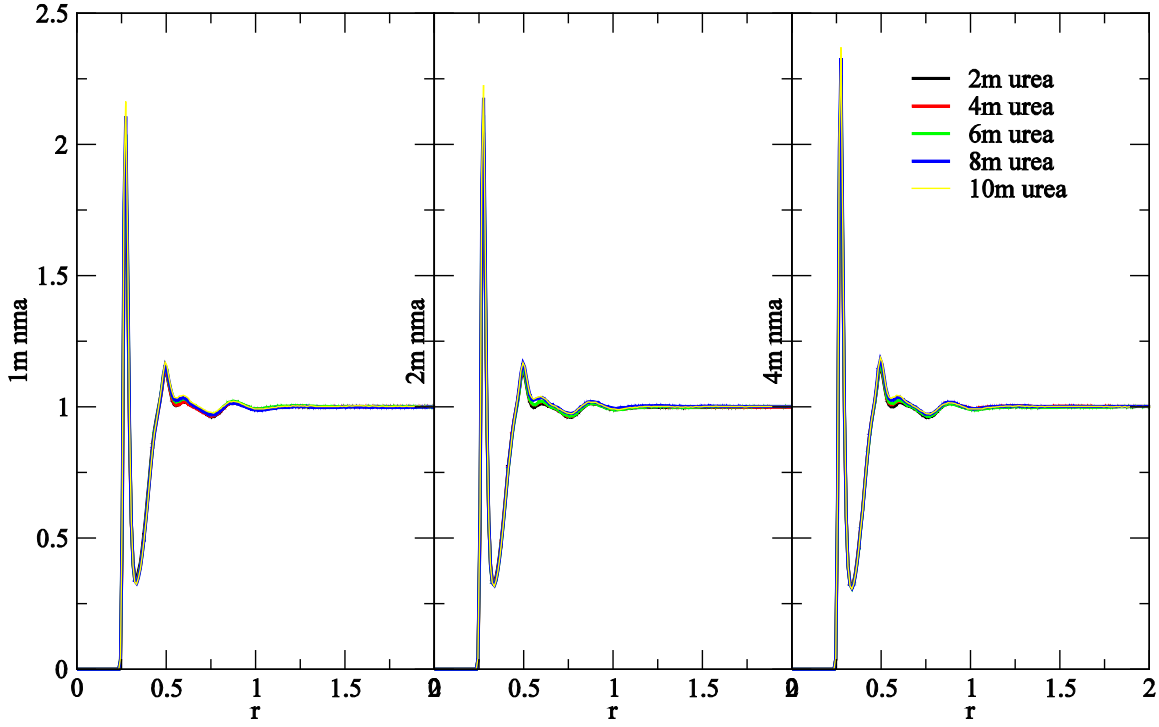
nmaO-Nurea



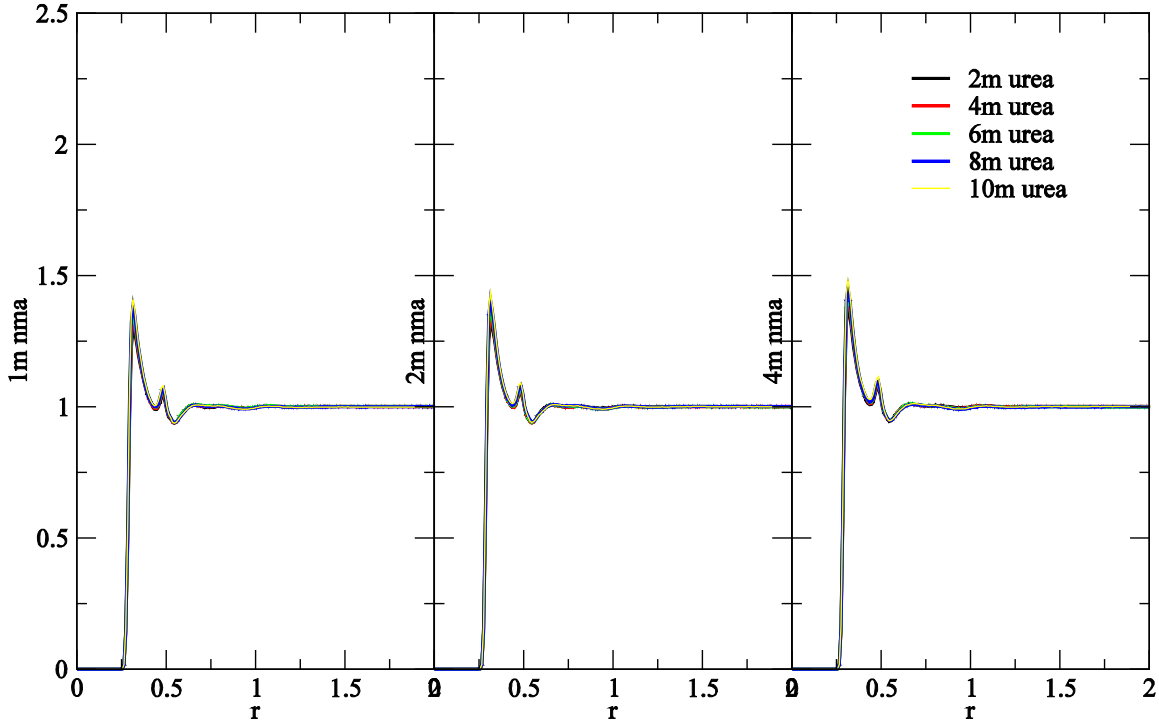
nmaN-Oh2o



nmaO-Oh2o



ureaN-Oh2o



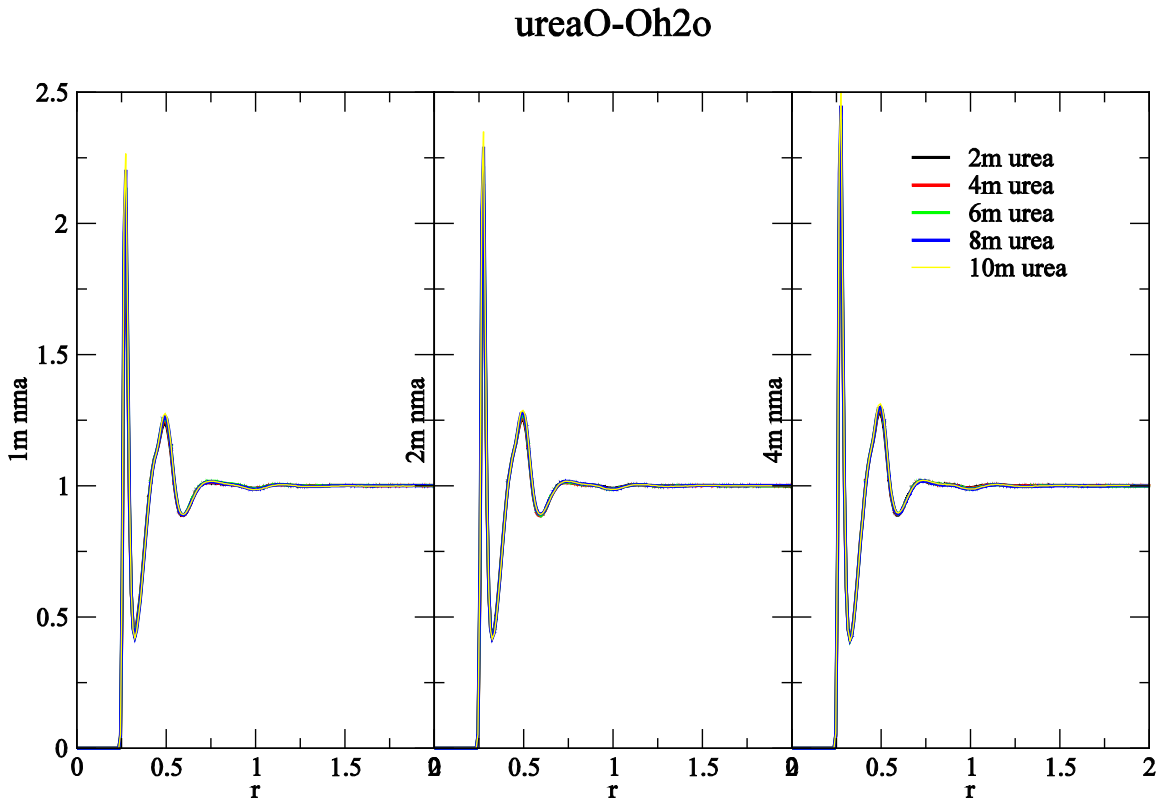
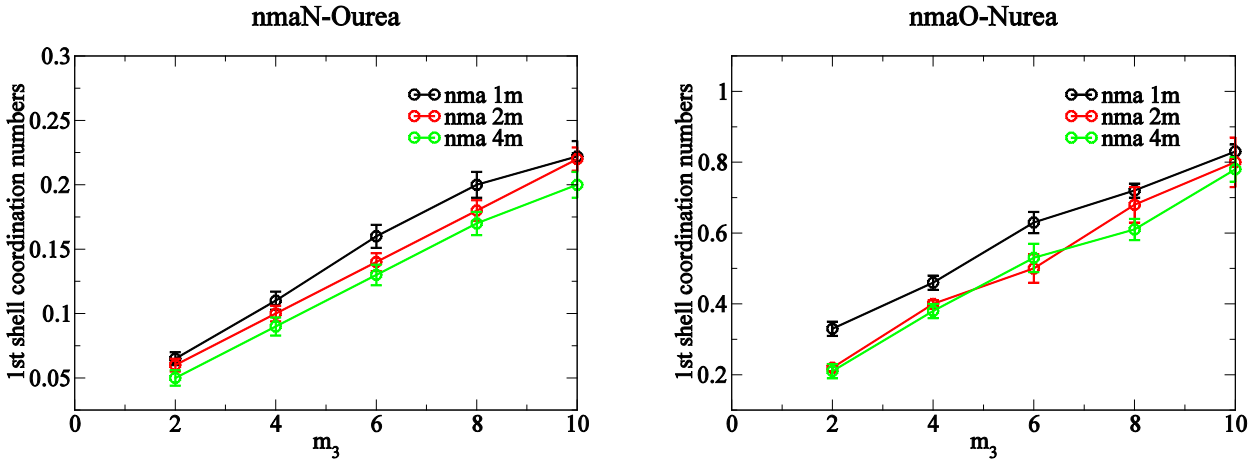
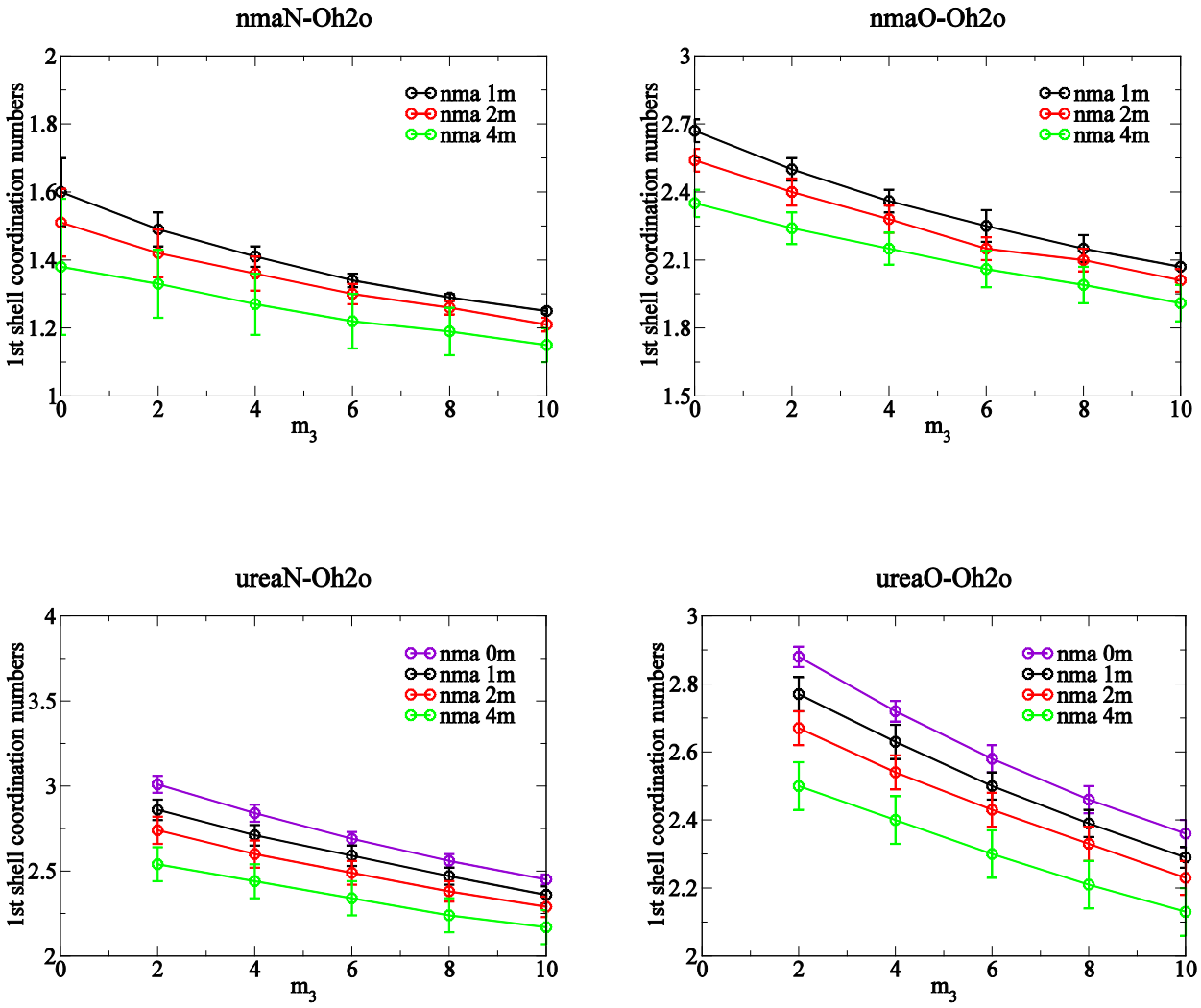


Figure 3.14 The simulated first shell coordination numbers from atom-atom radial distribution functions (g_{ij}) as a function of urea molality at 300 K.

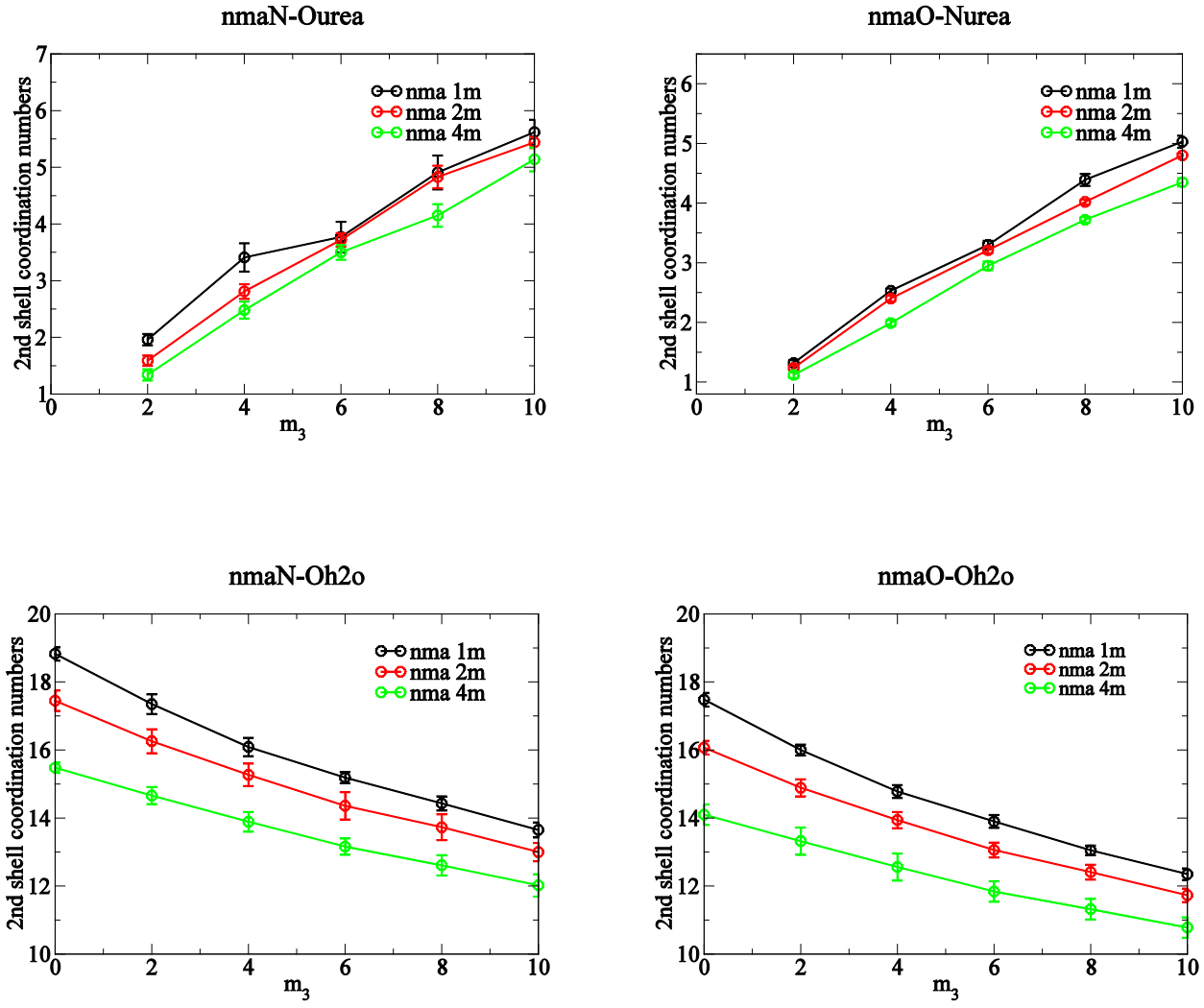


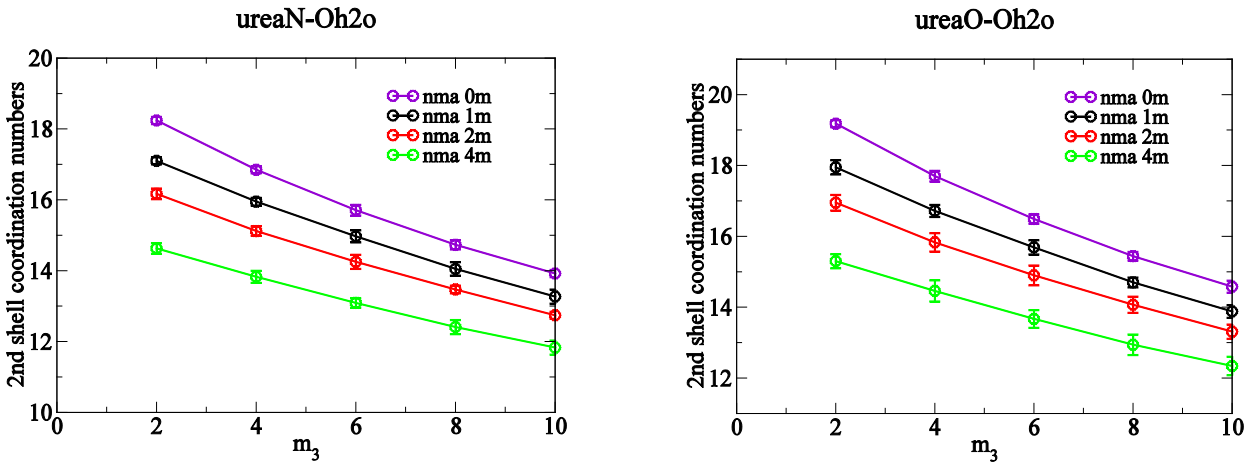


In order to more quantitatively study the change in molecular distribution with different compositions of NMA and urea, we calculated the coordination number of first and second shell from each of the atom-atom radial distribution functions. As shown in Figure 3.14, the first shell coordination numbers between NMA and urea increased as urea molality increased, indicating hydrogen bonding interactions increased between the solute (NMA) and cosolvent (urea). The coordination numbers with water went down as the urea concentration increases. All of the interactions decreased slightly as NMA increased, which means the addition of solute (NMA) actually would decrease the total number of H-bond between different species.

Figure 3.15 provides the second shell coordination numbers. They all have exactly the same trends as the first shell coordination numbers (increasing between NMA and urea as NMA decreased and urea increased, and all decreasing when interacted with water as NMA and urea increased). We can conclude that the addition of solute or cosolvent would promote the clustering of water molecules, while the interaction between urea and NMA was more complicated than expected.

Figure 3.15 The simulated second shell coordination numbers from atom-atom radial distribution functions (g_{ij}) as a function of urea molality at 300 K.

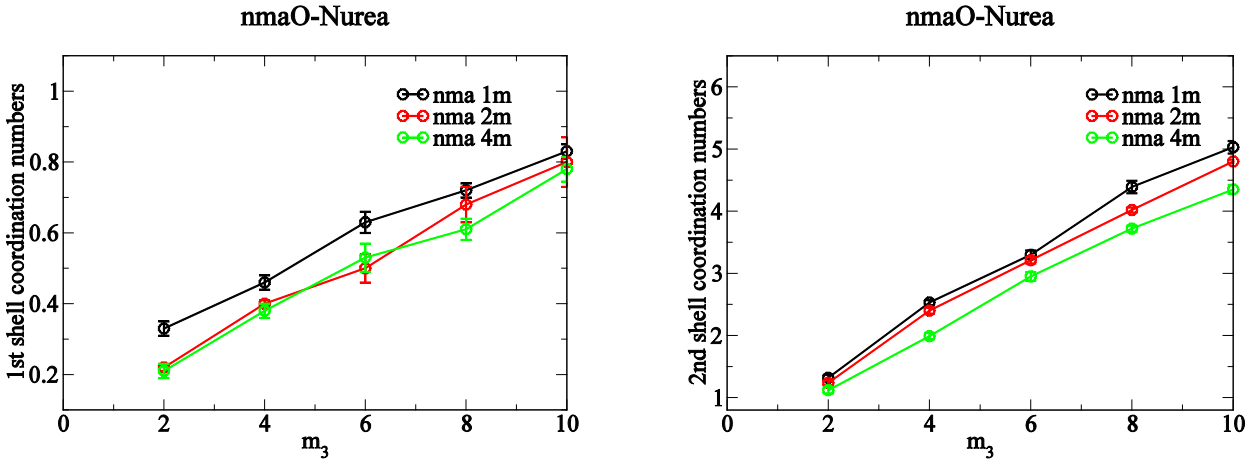


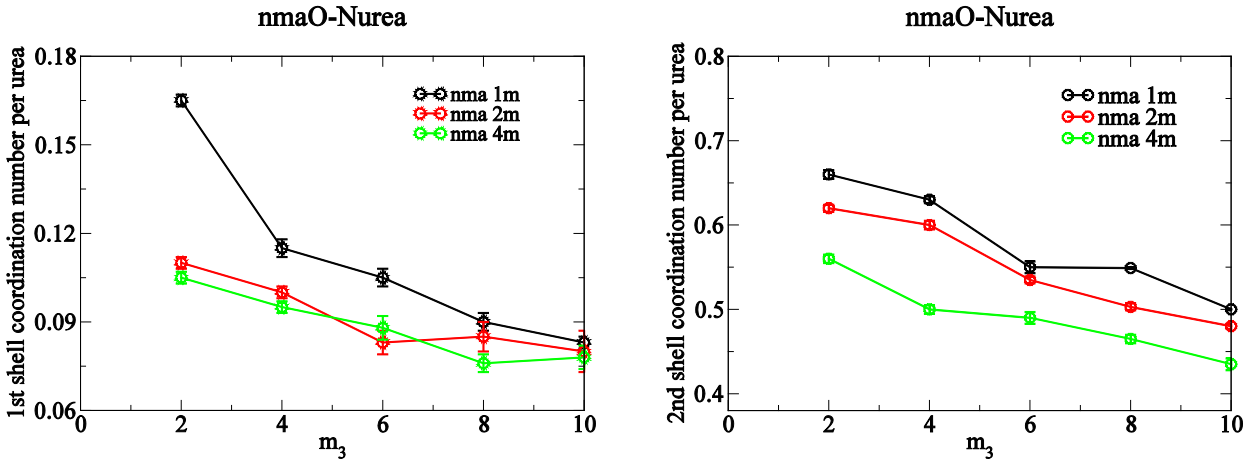


3.5.7 New insights

To further analyze the interaction between NMA and urea, as described by the calculated coordination numbers, we divided the above results by the corresponding urea molality to examine the contribution by each urea unit in the system.

Figure 3.16 First and second shell coordination numbers as a function of urea molality.



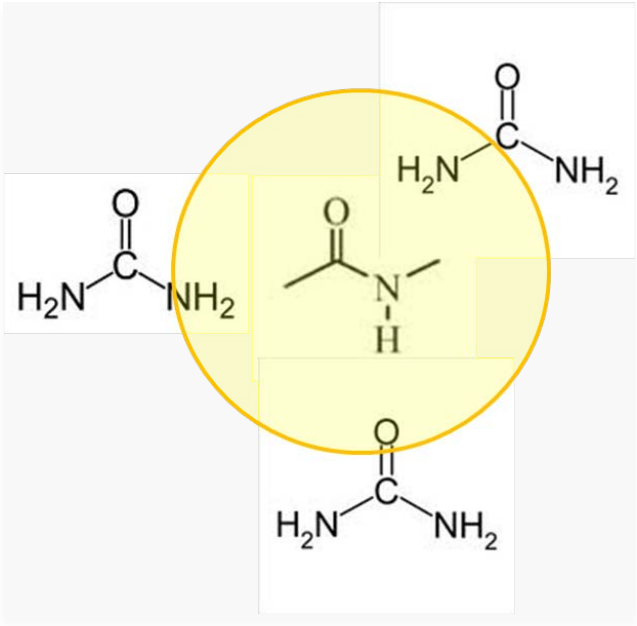


From Figure 3.16 it is clear that, even though the coordination numbers of first and second shell increased as the molality of urea increased, the average coordination number per urea actually decreased. In another word, the overall attraction between NMA and urea increased but not as much as the urea ration increased. Specifically, from the first shell coordination number the number of H-bond formed between urea and NMA went down in each molality of urea cosolvent as the molality of urea increased. This result suggested that less solute and more cosolvent could boost the formation of H-bonds between these two species, where the hydrogen bonding between solute and cosolvent might be helpful for protein denaturation, but not strongly related to the concentration change.

Another interesting discovery came from the comparison of center-of-mass and atom-atom radial distribution functions. The center of mass rdf's first shell between NMA and urea is at exactly the same distance to observe second shell of atom-atom nmaN-Ourea rdf with similar trends. This result provides some suggestions. Firstly, the center of mass rdf's first shell contained both the first and second shell of atom-atom rdf. It is the orientation of hydrogen that varied the distance between NMA and urea thus being separated into two shells of atom-atom rdf, but the center of mass was indeed located at first shell all the time. As shown in Figure 3.17, the urea at bottom could be counted as first shell nmaN-Ourea atom-atom rdf since the O from urea is heading to the N of NMA, the urea at top right could be counted as first shell nmaO-Nurea atom-atom rdf since the N from urea is heading to the O of NMA, while the urea at left

would be counted in neither first shell of atom-atom rdfs. But all these three urea could be counted in first shell of center-of-mass rdf.

Figure 3.17 A simple illustration of the orientation of urea molecules around NMA.



Secondly, based on the previous explanation, a large amount urea in the first shell center of mass rdf could be divided into second shell atom-atom rdf. Therefore, the rest of center-of-mass first shell, which corresponded to a small coordination number, belonged to the first shell of atom-atom rdf and was defined as the real H-bond. From statistic point of view, as more urea molecules around NMA molecule, there should be more chance that random arranged urea molecules could form H-bond to NMA. This explained that the first shell of atom-atom rdf increased as urea increased, but the peak of first shell is much smaller than center-of-mass rdf.

Thirdly, the second shell is much larger than first shell of atom-atom rdf, while its peak is at the same place where first peak of center-of-mass rdf formed. This tells us that all the urea that came from first shell center-of mass rdf and were not counted into first shell atom-atom rdf were locating in second shell of atom-atom rdf, and it included the majority of center-of-mass first shell. Besides, all the chemical potentials and preferential bindings we calculated before are

derived from KB integrals coming with center-of-mass rdfs. By comparing Figure 3.4 and Figure 3.7 ($\Gamma_{23} / \rho_3 = G_{23} - G_{21}$), we can prove the first shell of center-of-mass rdf was dominant and determined the trend. It is the whole set of molecules in center-of-mass rdf rather than the molecules in atom-atom first shell, that determined urea's nature as denaturant. So we can neglect the effect of molecules in second shell of atom-atom rdf. In fact, we believe the molecules in first shell center-of-mass rdf influenced the denaturation procedure as well.

3.6 Conclusions

In summary, NMA is more favorable in urea solutions than in pure water. The explicit expressions derived from the KB integrals for three component solution mixtures helped to clarify the relationships like chemical potentials, preferential binding parameters and provide a clear picture of the effects of cosolvents so that reasonable transfer free energies are obtained from the simulations. In addition the second contribution ($-\Gamma_{23}\ \mu_{33}$) to the total transfer free energy is dominant.

According to KB Analysis based on center-of-mass rdf, the $G_{23} - G_{21}$ decreased as the urea concentration increased, following the changes in the first shell of center-of-mass rdfs. However, the atom-atom first shell of NMA and urea that could explain formation of H-bond of urea and NMA, is unchanged (no change in H-bonding) compared to the second shell. And we found that H-bonding between urea and NMA is weak at best.

By comparing center-of-mass rdf and atom-atom rdf in first and second shell, our KB integrals indicated the loss of urea from center-of-mass first shell, which didn't contribute to H-bonds, transferred to the atom-atom second shell and to some degree influenced denaturation. The results suggest a role for packing effects rather than simple changes in hydrogen bonding, therefore the protein denaturation doesn't strongly depend on urea concentration.

3.7 References

(1) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.

(2) Timasheff, S. N. *Advances in Protein Chemistry, Vol 51* **1998**, *51*, 355.

(3) Vonhippe.Ph; Schleich, T. *Accounts of Chemical Research* **1969**, *2*, 257.

(4) Franks, F.; Eagland, D. *CRC Crit Rev Biochem* **1975**, *3*, 165.

(5) Anderson, C. F.; Record, M. T. *Annual Review of Biophysics and Biophysical Chemistry* **1990**, *19*, 423.

(6) Timasheff, S. N. *Biochemistry* **1992**, *31*, 9857.

(7) Buck, M. *Q Rev Biophys* **1998**, *31*, 297.

(8) Collins, K. D.; Washabaugh, M. W. *Q Rev Biophys* **1985**, *18*, 323.

(9) Nosworthy, N. J.; Ginsburg, A. *Protein Science* **1997**, *6*, 2617.

(10) Voziyan, P. A.; Jadhav, L.; Fisher, M. T. *Journal of Pharmaceutical Sciences* **2000**, *89*, 1036.

(11) Baynes, B. M.; Trout, B. L. *Biophysical Journal* **2004**, *87*, 1631.

(12) Klimov, D. K.; Straub, J. E.; Thirumalai, D. *Proc Natl Acad Sci U S A* **2004**, *101*, 14760.

(13) Liu, R.; Barkhordarian, H.; Emadi, S.; Park, C. B.; Sierks, M. R. *Neurobiology of Disease* **2005**, *20*, 74.

(14) Narayanan, S.; Reif, B. *Biochemistry* **2005**, *44*, 1444.

(15) Smith, P. E. *Biophysical Chemistry* **2005**, *113*, 299.

(16) Arakawa, T.; Ejima, D.; Tsumoto, K.; Obeyama, N.; Tanaka, Y.; Kita, Y.; Timasheff, S. N. *Biophysical Chemistry* **2007**, *127*, 1.

(17) Throop, G. J.; Bearman, R. J. *Journal of Chemical Physics* **1966**, *44*, 1423.

(18) Levitt, M.; Park, B. H. *Structure* **1993**, *1*, 223.

(19) Otting, G. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1997**, *31*, 259.

(20) Otting, G. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1998**, *32*, 191.

(21) Otting, G.; Liepinsh, E.; Wuthrich, K. *Science* **1991**, *254*, 974.

(22) Liepinsh, E.; Otting, G. *Nat Biotechnol* **1997**, *15*, 264.

(23) Ponstingl, H.; Otting, G. *Journal of Biomolecular Nmr* **1997**, *9*, 441.

(24) Liepinsh, E.; Sodano, P.; Tassin, S.; Marion, D.; Vovelle, F.; Otting, G. *Journal of Biomolecular Nmr* **1999**, *15*, 213.

(25) Kuntz, I. D., Jr.; Kauzmann, W. *Adv Protein Chem* **1974**, *28*, 239.

(26) Pessen, H.; Kumosinski, T. F. *Methods in Enzymology* **1985**, *117*, 219.

(27) Rupley, J. A.; Careri, G. *Adv Protein Chem* **1991**, *41*, 37.

(28) Svergun, D. I.; Richard, S.; Koch, M. H. J.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proc Natl Acad Sci U S A* **1998**, *95*, 2267.

(29) Robinson, D. R.; Jencks, W. P. *J Am Chem Soc* **1965**, *87*, 2462.

(30) Nandi, P. K.; Robinson, D. R. *Biochemistry* **1984**, *23*, 6661.

(31) Gill, S. J.; Downing, M.; Hutson, J.; Clopton, J. R. *Journal of Physical Chemistry* **1961**, *65*, 1432.

(32) Wetlaufer, D. B.; Coffin, R. L.; Malik, S. K.; Stoller, L. *Journal of the American Chemical Society* **1964**, *86*, 508.

(33) Schonert, H.; Stroth, L. *Biopolymers* **1981**, *20*, 817.

(34) Kresheck, G. C.; Benjamin, L. *Journal of Physical Chemistry* **1964**, *68*, 2476.

(35) Savage, J. J.; Wood, R. H. *Journal of Solution Chemistry* **1976**, *5*, 733.

(36) Sijpkes, A. H.; Somsen, G.; Lilley, T. H. *Journal of the Chemical Society-Faraday Transactions* **1990**, *86*, 2943.

(37) Brandts, J. F.; Hunt, L. *J Am Chem Soc* **1967**, *89*, 4826.

(38) Johnson, C. M.; Fersht, A. R. *Biochemistry* **1995**, *34*, 6795.

(39) Makhatadze, G. I.; Privalov, P. L. *J Mol Biol* **1992**, *226*, 491.

(40) Neri, D.; Billeter, M.; Wider, G.; Wuthrich, K. *Science* **1992**, *257*, 1559.

(41) Pace, C. N.; Laurents, D. V.; Thomson, J. A. *Biochemistry* **1990**, *29*, 2564.

(42) Pace, C. N.; Laurents, D. V.; Erickson, R. E. *Biochemistry* **1992**, *31*, 2728.

(43) Pazandrade, M. I.; Jones, M. N.; Skinner, H. A. *European Journal of Biochemistry* **1976**, *66*, 127.

(44) Schellman, J. A.; Gassner, N. C. *Biophys Chem* **1996**, *59*, 259.

(45) DeKoster, G. T.; Robertson, A. D. *Biophysical Chemistry* **1997**, *64*, 59.

(46) Makhatadze, G. I.; Privalov, P. L. *Journal of Molecular Biology* **1992**, *226*, 491.

(47) Hua, L.; Zhou, R. H.; Thirumalai, D.; Berne, B. J. *Proc Natl Acad Sci U S A* **2008**, *105*, 16928.

(48) Wallqvist, A.; Covell, D. G.; Thirumalai, D. *Journal of the American Chemical Society* **1998**, *120*, 427.

(49) Rezus, Y. L. A.; Bakker, H. J. *Proc Natl Acad Sci U S A* **2006**, *103*, 18417.

(50) Tanford, C. *Journal of the American Chemical Society* **1964**, *86*, 2050.

(51) Bennion, B. J.; Daggett, V. *Proc Natl Acad Sci U S A* **2003**, *100*, 5142.

(52) Finer, E. G.; Franks, F.; Tait, M. J. *Journal of the American Chemical Society* **1972**, *94*, 4424.

(53) Tanford, C. *Journal of the American Chemical Society* **1962**, *84*, 4240.

(54) Nozaki, Y.; Tanford, C. *Journal of Biological Chemistry* **1963**, *238*, 4074.

(55) Tanford, C. *Adv Protein Chem* **1968**, *23*, 121.

(56) Arakawa, T.; Bhat, R.; Timasheff, S. N. *Biochemistry* **1990**, *29*, 1914.

(57) Arakawa, T.; Timasheff, S. N. *Biochemistry* **1982**, *21*, 6545.

(58) Timasheff, S. N. *Proc Natl Acad Sci U S A* **2002**, *99*, 9721.

(59) Arakawa, T.; Timasheff, S. N. *Archives of Biochemistry and Biophysics* **1983**, *224*, 169.

(60) Bolen, D. W.; Rose, G. D. *Annual Review of Biochemistry* **2008**, *77*, 339.

(61) Auton, M.; Bolen, D. W. *Osmosensing and Osmosignaling* **2007**, *428*, 397.

(62) Auton, M.; Bolen, D. W. *Biochemistry* **2004**, *43*, 1329.

(63) Zou, Q.; Habermann-Rottinghaus, S. M.; Murphy, K. P. *Proteins-Structure Function and Genetics* **1998**, *31*, 107.

(64) Zou, Q.; Bennion, B. J.; Daggett, V.; Murphy, K. P. *Journal of the American Chemical Society* **2002**, *124*, 1192.

(65) Hong, J.; Capp, M. W.; Anderson, C. E.; Record, M. T. *Biophysical Chemistry* **2003**, *105*, 517.

(66) Zhang, W. T.; Capp, M. W.; Bond, J. P.; Anderson, C. F.; Record, M. T. *Biochemistry* **1996**, *35*, 10506.

(67) Courtenay, E. S.; Capp, M. W.; Record, M. T. *Protein Science* **2001**, *10*, 2485.

(68) Courtenay, E. S.; Capp, M. W.; Saecker, R. M.; Record, M. T. *Proteins-Structure Function and Genetics* **2000**, 72.

(69) Hong, J.; Capp, M. W.; Anderson, C. F.; Saecker, R. M.; Felitsky, D. J.; Anderson, M. W.; Record, M. T. *Biochemistry* **2004**, *43*, 14744.

(70) Felitsky, D. J.; Record, M. T. *Biochemistry* **2004**, *43*, 9276.

(71) Felitsky, D. J.; Record, M. T. *Biochemistry* **2003**, *42*, 2202.

(72) Hong, J.; Capp, M. W.; Saecker, R. M.; Record, M. T. *Biochemistry* **2005**, *44*, 16896.

(73) Cannon, J. G.; Anderson, C. F.; Record, M. T. *Journal of Physical Chemistry B* **2007**, *111*, 9675.

(74) Guinn, E. J.; Pegram, L. M.; Capp, M. W.; Pollock, M. N.; Record, M. T. *Proc Natl Acad Sci U S A* **2011**, *108*, 16932.

(75) Shimizu, S.; Chan, H. S. *Proteins-Structure Function and Genetics* **2002**, *49*, 560.

(76) Shimizu, S.; Boon, C. L. *Journal of Chemical Physics* **2005**, *122*, 049901.

(77) Shimizu, S.; Boon, C. L. *Journal of Chemical Physics* **2004**, *121*, 9147.

(78) Shimizu, S. *Journal of Chemical Physics* **2005**, *122*, 199901.

(79) Shimizu, S. *Journal of Chemical Physics* **2004**, *120*, 4989.

(80) Shimizu, S.; Matubayasi, N. *Chemical Physics Letters* **2006**, *420*, 518.

(81) Shimizu, S. *Chemical Physics Letters* **2011**, *517*, 76.

(82) O'Brien, E. P.; Dima, R. I.; Brooks, B.; Thirumalai, D. *Journal of the American Chemical Society* **2007**, *129*, 7346.

(83) Mountain, R. D.; Thirumalai, D. *Journal of the American Chemical Society* **2003**, *125*, 1950.

(84) Tobi, D.; Elber, R.; Thirumalai, D. *Biopolymers* **2003**, *68*, 359.

(85) Zangi, R.; Zhou, R. H.; Berne, B. J. *Journal of the American Chemical Society* **2009**, *131*, 1535.

(86) Canchi, D. R.; Paschek, D.; Garcia, A. E. *Journal of the American Chemical Society* **2010**, *132*, 2338.

(87) Stumpe, M. C.; Grubmuller, H. *Journal of Physical Chemistry B* **2007**, *111*, 6220.

(88) Stumpe, M. C.; Grubmuller, H. *Journal of the American Chemical Society* **2007**, *129*, 16126.

(89) Stumpe, M. C.; Grubmuller, H. *Biophysical Journal* **2009**, *96*, 3744.

(90) Kirkwood, J. G.; Buff, F. P. *Journal of Chemical Physics* **1951**, *19*, 774.

(91) Kang, M.; Smith, P. E. *Journal of Computational Chemistry* **2006**, *27*, 1477.

(92) Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochemistry and Biophysics* **2008**, *50*, 1.

(93) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2003**, *107*, 3891.

(94) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2005**, *109*, 15080.

(95) Ben-Naim, A. *Statistical thermodynamics for chemists and biochemists.*; Plenum Press: New York, **1992**.

(96) Schellman, J. A. *Biophysical Chemistry* **2002**, *96*, 91.

(97) Eisenberg, H. *Biological macromolecules and polyelectrolytes in solution*; Clarendon Press: Oxford, **1976**.

(98) Wyman, J. *Adv Protein Chem* **1964**, *19*, 223.

(99) Parsegian, V. A.; Rand, R. P.; Fuller, N. L.; Rau, D. C. *Methods in Enzymology* **1986**, *127*, 400.

(100) Kirkwood, J. G.; Goldberg, R. J. *Journal of Chemical Physics* **1950**, *18*, 54.

(101) Eisenberg, H. *Biophysical Chemistry* **1994**, *53*, 57.

(102) Stockmayer, W. H. *Journal of Chemical Physics* **1950**, *18*, 58.

(103) Casassa, E. F.; Eisenberg, H. *Adv Protein Chem* **1964**, *19*, 287.

(104) Schellman, J. A. *Biophysical Chemistry* **1990**, *37*, 121.

(105) Timasheff, S. N. *Annual Review of Biophysics and Biomolecular Structure* **1993**, *22*, 67.

(106) Kang, M.; Smith, P. E. *Journal of Chemical Physics* **2008**, *128*.

(107) Kang, M.; Smith, P. E. *Fluid Phase Equilibria* **2007**, *256*, 14.

(108) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *Journal of Physical Chemistry* **1987**, *91*, 6269.

(109) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *Journal of Chemical Theory and Computation* **2008**, *4*, 435.

(110) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684.

(111) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *Journal of Computational Chemistry* **1997**, *18*, 1463.

(112) Miyamoto, S.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 952.

(113) Darden, T.; York, D.; Pedersen, L. *Journal of Chemical Physics* **1993**, *98*, 10089.

(114) Korolev, V. P. *Journal of Structural Chemistry* **2008**, *49*, 660.

Chapter 4 - Development and Testing of Torsional Potentials for Peptides and Proteins

4.1 Abstract

Recently, we have been developing a series of force fields for the eventual simulation of biological systems by attempting to accurately reproduce the experimental Kirkwood-Buff (KB) integrals observed in solution mixtures. The resulting KB derived force fields (KBFF) provide quantitative data concerning solute-solute, solute-solvent and solvent-solvent distributions in solution. Here, we describe our latest results to obtain a complete balanced force field for peptides and proteins. The general approach will be outlined and simulations of a variety of peptides and proteins in aqueous solutions will be presented to demonstrate the reliability of the force field.

4.2 Introduction

Simulation studies of proteins and peptides need to be performed by accurate empirical force fields.[1,2] A number of force fields are currently available, including CHARMM19[3] and 22,[4] OPLS,[5] AMBER,[6] and GROMOS,[7] which are all specifically designed for the study of biological systems. Although these force fields have been extensively used to study a wide variety of systems in some detail, they still possess several inherent shortcomings which reduce their accuracy. First, relates to the sampling problem. The simulation time should be long enough to enable the sampling of all relevant molecular conformations. This essentially determines the precision of the simulation results. Second, relates to the force field problem. An inaccurate energy function may bias the simulation towards incorrect behavior, thereby significantly affecting the accuracy of the data.[8]

Unlike small molecules, which may not be severely affected by sampling limitations, large molecules like proteins with many potential conformations are more susceptible to sampling problems. Many approaches to improve the degree of sampling in molecular simulations have been developed and include techniques relating to both software and hardware issues.[8-11] With current approaches and computers one can perform simulations of reasonably large systems on the microsecond timescale when applying enhanced sampling.[8-11] Unfortunately, many of these longer MD simulations have shown the defect from existing force

fields.[9,12] Thus continuous improvements in force field accuracy are still required. Some studies have focused on developing more accurate polarizable force fields,[13] but at the price of more expensive computer simulations. For this reason, many people believe that there is still room to improve existing non-polarizable force fields.

Many of the most notable efforts to improve the accuracy of current force fields is the re-parameterization of the torsional potentials. These potentials (see Equation 4.1) determine the conformational preferences (alpha versus beta) observed for the amino acid backbone and side chains,

$$\sum_{torsions} k_{\phi}[1 + \cos(n\phi - \phi_s)] \tag{4.1}$$

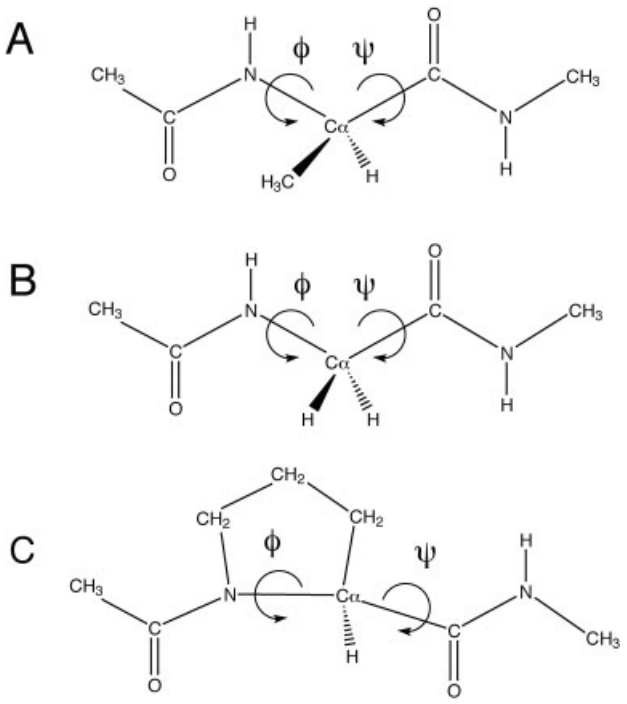
Modifications to the protein backbone and sidechain torsional potentials have been performed for both the Amber and CHARMM force fields and have led to significant improvements in accuracy.[14] For example, the Amber99SB force field,[15] and the CMAP correction for CHARMM22,[16] both focused on the backbone torsional potentials, while the recently developed Amber 99SB ILDN force field[17] included improved sidechain dihedral potentials.

In this study, we present an initial version of the complete Kirkwood Buff Force Field for proteins and peptides. Specifically, the development of initial backbone and side chain torsional potentials. This optimization effort includes the use of QM calculations to provide the required force field parameters. Unfortunately, even after obtaining very good reproduction of QM data, additional MD simulations of small peptides and proteins still indicated systematic errors for the φ, ψ potentials. Thus, a further refinement against the known alpha and beta preferences of small peptides was therefore performed to provide the final parameter set and hopefully remove any inherent biases in the force field.

4.2.1 Backbone Torsions

Many recent efforts to improve biomolecular force fields have centered around the critically important ϕ and ψ degrees of freedom. Model systems typically used to study this problem are the alanine, glycine and proline dipeptides, which are actually single blocked amino acids, ACE-AA-NMH, where AA represents amino acid Ala, Gly or Pro. As shown in Figure 4.1, there are two peptide bonds surrounding the central ϕ (C-N-C^{α}-C), ψ (N-C^{α}-C-N) dihedral angles in each dipeptide.

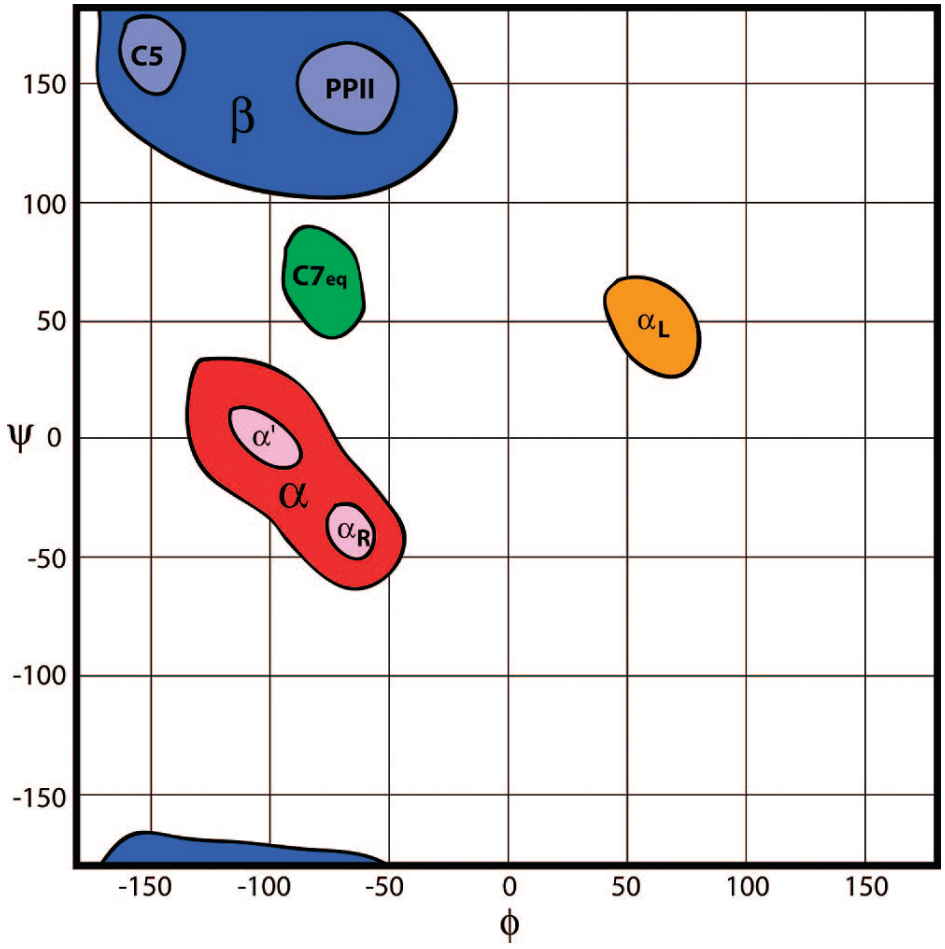
Figure 4.1 (A) blocked alanine dipeptide, (B) blocked glycine dipeptide and (C) blocked proline dipeptide (figure taken from Mackerell's paper).[16]



The distribution of the φ, ψ dihedral angles determines the major features of the well-known Ramachandran map.[18,19] Ramachandran maps provide two dimensional representations of the observed φ, ψ preferences of amino acids from which it is easy to recognize the common secondary structures such as α-helices and β-sheets. The original map was composed from a large set of experimental data such as crystallography and NMR, and is in qualitative agreement with the conformational energetics based on simple interaction potentials. All amino acids, except for glycine and proline, have similar backbone structures and therefore similar Ramachandran plots as shown in Figure 4.2. For non-glycine and non-proline residues, the α region has a minimum at ϕ = -60, ψ = -50, while the β region (ϕ = -60 to -170, ψ = 120 to 170) has two minima at both C5 (ϕ = -150, ψ = 150) and PPII (ϕ = -60, ψ = 140). It is common for the blocked alanine dipeptide to be used as a simple model to study non-glycine non-proline peptide backbones. Since the C^{α} atom in glycine is not chiral, one should additional symmetry patterns

for positive and negative values of ϕ in the Ramachandran plot. The conformations of proline residues are restricted to the α_R, PPII, and $C7_{eq}$ regions.

Figure 4.2 Schematic overview of the major conformational basins sampled by φ/Ψ backbone torsion angles in nonglycine, nonproline peptide residues (figure taken from Feig's paper).[20]



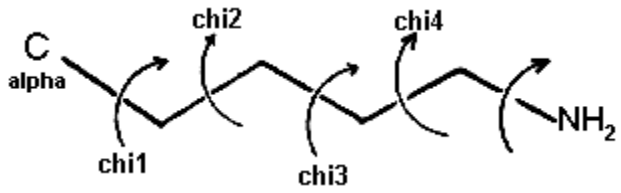
In most theoretical and simulation studies of proteins the conformational preferences and energies of the peptide backbone are directly related to the force field torsional parameters. Hence, many quantum mechanical (QM) and molecular mechanics (MM) studies have been performed on the alanine, glycine and proline dipeptides,[21-32] in an effort to fully understand and reproduce these preferences. In fact, most attempts to derive or refine torsional parameters start

with the study of the dipeptides and the goal of fitting to the QM potential energy surfaces. Unfortunately, it should be remembered that these are gas phase potential energy surfaces. Due to polarization and solvation effects the free energy surface in solution can be quite different.[16,27,33] Therefore, in the present study a slightly different approach was used during the fitting procedure compared with other studies. Specifically, torsional potentials were assumed to be transferable from the gas phase to solution. Since the gas phase charge distribution is different from the condensed phase charge distribution, we simply adopted a scaling factor for our condensed phase charges,[34] in order to mimic the gas phase charge distribution during the fitting procedure. The fitted surfaces for the gas phase are then compared with *ab initio* gas phase surfaces. This appears to produce very reasonable potentials. However, in subsequent simulations of small peptides there remained a bias toward alpha conformations. Hence, we developed additional small corrections terms for φ, ψ dihedral parameters in order to better reproduce the experimental NMR data.

4.2.2 Sidechain Torsions

As shown in Figure 4.3, the amino acid has side chain has freedom for rotation as well. The standard dihedral notation assigns rotation around Cα and Cβ as defined by χ1, with rotation around Cβ and Cγ defined by χ2, etc.

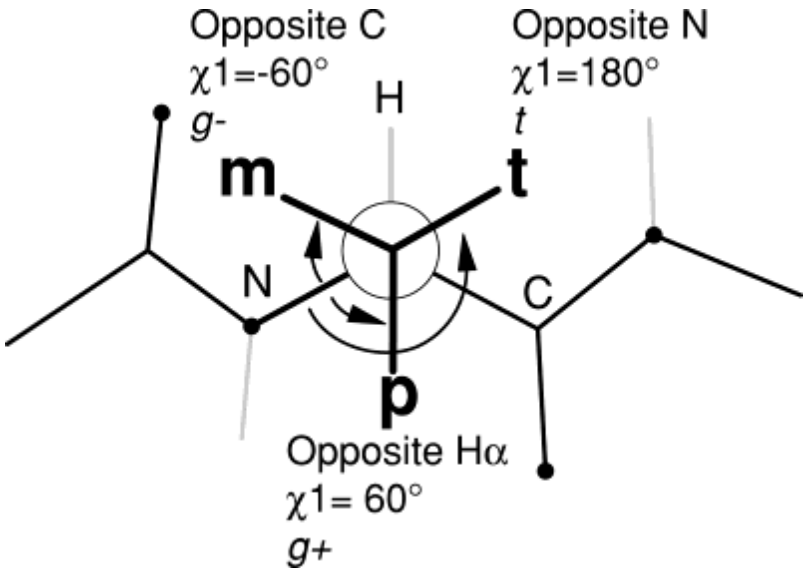
Figure 4.3 The dihedrals χ1, χ2, χ3, and χ4 for the sidechain of Lysine.



The rotation of the χ angles, especially the χ1 angle, is highly dependent on the backbone φ, ψ dihedral distribution. As shown in Figure 4.4, when the backbone dihedrals are determined, χ1 angle only have three possible conformations, which are g^- (χ1 = -60°), g^+ (χ1 = 60°) and t (χ1 = 180°) respectively. Therefore, given correct backbone φ, ψ dihedral distribution, it would be not difficult to determine χ1 dihedral distribution based on available experimental data.

In developing the Kirkwood Buff force field we require χ torsional potentials for the amino acid side chains. Among the torsional degrees of freedom in proteins and peptides, the χ_1 torsions are expected to be second only to the backbone in importance for describing conformational energetics. Initial attempts to provide side chain parameters also used available QM calculations concerning relation conformational energies and barriers to rotation.[35] However, further refinement of the side chain potentials was required. We have therefore focused our efforts on correcting the sidechain distribution for the condensed phase and used a three step procedure to develop the force field parameters. First, we generally parameterized the various different types of torsion based on atom type, C-C-C-C or C-C-C-O, etc by fitting gas phase QM data. In doing so we were careful to scale the potentials down in cases where there were multiple heavy atom definitions for the torsion. Second, we identified problematic residues with large deviations from the expected database behavior for χ1 populations for biological systems. Third, we used additional potentials for the χ1 dihedrals by fitting to the residue specific distribution of χ_1 dihedrals observed in the Protein Data Bank (condensed phase).[36]

Figure 4.4 The illustration of the relationships for the side-chain dihedral angle χ1. Each of the three staggered conformations is measured from the backbone N (figure taken from Lovell's paper).[33]



4.3 Methods

4.3.1 Backbone Torsions

The nonbonded parameters used for the glycine, alanine and proline dipeptides are presented in Table 4.1 and were taken from our previous study of aqueous amides.[34] The bonded parameters, with the exception of the torsional potentials, were taken (with permission) from the GROMOS G53a6 force field. Development of the torsional parameters was then performed in a stepwise manner. First, the gas phase energy surfaces were generated for the glycine and alanine dipeptides in the absence of any torsional potentials and using a series of different scaling factors for the atomic charges and 1-4 LJ and Coulomb interactions. All bonds were constrained using Lincs and no cutoffs were used. The scaling factors which produced the best agreement with the gas phase QM maps (15º resolution) were then adopted for the remaining fitting procedure. These factors were 0.1 and 0.5 for the 1-4 LJ and Coulomb interactions, respectively. The optimum scaling factor for the overall charge distribution was $0.7^{½}$ = 0.837, and was implemented by using a relative permittivity of 1/0.7 = 1.43 in the calculations. Second, the torsional potentials for φ and ψ were generated by fitting to the QM glycine dipeptide map using the above scaling factors. These potentials were then used for the φ and ψ torsions of all residues. Third, the φ′ (C -N-C^{α}-C^{β}) and ψ′ (N -C-C^{α}-C^{β}) potentials were then generated by fitting to the alanine dipeptide QM map.[16] When fitting to the QM maps a series of biases were applied to ensure a good fit in the regions of most importance. Hence, the weights associated with the points immediately surrounding (±15º) φ/ψ values of 60/45, 60/ -150, 90/-75, 150/-150 and their symmetry related regions were increased by a factor of 10 for the glycine dipeptide, while the weights of the points immediately surrounding φ/ψ values of -60/-45, -60/150, -90/75, -150/150 and 60/45 were increased by a factor of 10 for the alanine dipeptide. The torsional potentials for rotation around ω were generated to reproduce the experimental data. The torsional potentials involving the proline side chain atoms were optimized to produce reasonable agreement with experimental pseudo rotation rates (P. E. Smith, unpublished data).

The nonbonded energy is described by a Lennard-Jones 6-12 plus Coulomb potential with geometric combination rules for both ε_{ij} and σ_{ij}. The LJ nonbonded 1-4 interactions were scaled by a factor of 0.1, while the Coulomb 1-4 interactions were scaled by a factor of 0.5.

Table 4.1 The KBFF non-bonded parameters.

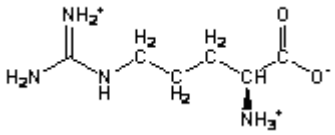
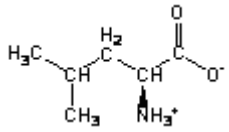
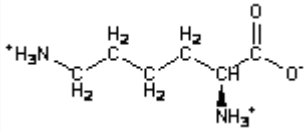
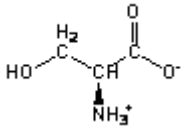
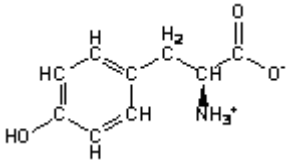
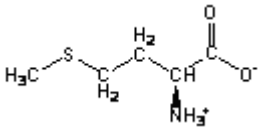
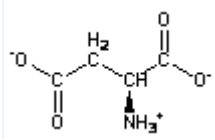
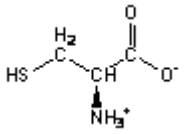
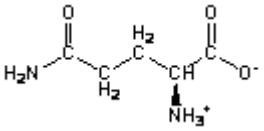
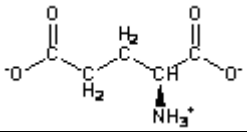
Atom	ε_{ii} kJ/mol	σ_{ii} nm	q |e|
C, C^{δ} (gln)	0.330	0.336	0.62
O, $O^{\varepsilon 1}$ (gln)	0.560	0.310	-0.62
N, $N^{\varepsilon 2}$ (gln)	0.500	0.311	-0.70
H	0.088	0.158	0.36
$H^{\varepsilon 2}$ (gln)	0.088	0.158	0.35
CH3(C)	0.869	0.375	0.0
CH3(N)	0.869	0.375	0.34
C^{α} (gly)	0.411	0.407	0.34
C^{α} (ala, gln)	0.095	0.502	0.34
C^{α} (pro)	0.095	0.502	0.35
C^{β} (ala)	0.869	0.375	0.0
C^{β}, C^{γ} (pro, gln)	0.411	0.407	0.0
C^{δ} (pro)	0.411	0.407	0.35
O (SPC/E)	0.6506	0.3166	-0.8476
H (SPC/E)	0.0	0.0	0.4238

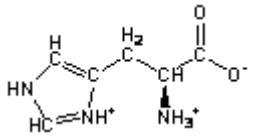
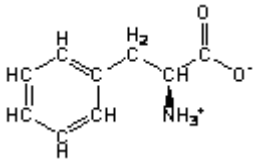
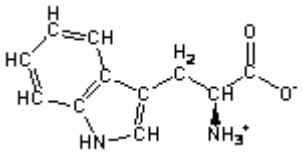
4.3.2 Sidechain Torsions

The torsional potentials used for sidechain such as C-C-C-C, C-C-C-N, C-C-O-H etc, (see Table 4.2) were optimized in this section. The torsional energy profiles were initially generated to mimic gas phase results by comparing with QM potential calculations from ab initio calculations at the RHF/6-31G* level.[37] Further improvement of the potentials for use in the solution phase was dependent on the atoms involved in the torsional degree of freedom. For torsions containing only non-polar atoms it was assumed their behavior should be similar in both the gas phase and condensed phase, and therefore there was no need for further refinement. For torsions involving polar atoms the resulting condensed phase distributions were examined carefully and modifications were made if necessary – when the distributions differed from the

PDB distributions. Finally, additional χ1 dihedral terms were developed for each amino acid to match the PDB distributions.

Table 4.2 Examples of dihedrals in amino acid side chains.

Dihedral	Amino Acid	Structure
Cγ-Cδ-Nε-Cζ	Arg	
Cα-Cβ-Cγ-Cδ	Leu	
Cγ-Cδ-Cε-Nζ	Lys	
Cα-Cβ-Oγ-Hδ	Ser	
Cε-Cζ-Oη-Hθ	Tyr	
Cβ-Cγ-Sδ-Cε	Met	
Oδ´-Cγ-Oδ-Hε	Asp	
Cα-Cβ-Sγ-Hδ	Cys	
Cβ-Cγ-Cδ-Nε	Gln	
Cβ-Cγ-Cδ-Oε	Glu	

Cβ-Cγ-Nδ-Hε	His	
Cα-Cβ-Cγ-Cδ	Phe	
Cα-Cβ-Cγ-Cδ	Trp	

The condensed phase distributions for the KBFF models were determined from simulations of the appropriate dipeptide. A solvated terminally blocked amino acid sequence ACE-Ala-X-Ala-NHM, where X is any one of the twenty natural amino acids, was simulated in a cubic box with sides of ~40 Å containing ~400 water molecules. Protonation states were chosen to correspond to neutral pH with an additional sodium or chloride ion added to the system to achieve neutrality. Each system was then equilibrated at 300K and 1atm with 1ns of MD simulation in the NPT ensemble. REMD simulations[38,39] were then performed in the NVT ensemble for 50 ns at a series of temperatures ranging from 300K to 650K. Temperature exchanges were attempted every 2 ps. The resulting population distributions were then used to determine the percentage population of the p, m, and t wells.

4.3.3 Small Peptides

A series of small peptides with well characterized secondary structure were then selected to provide some validation of the Kirkwood Buff Force Field. The full list of peptides is shown in Table 4.3. MD simulations were performed using the KBFF models. Simulation parameters were the same as in the simulations of amino acid side chain (see Section 4.4). Each peptide was initially solvated in 60 Å cubic water boxes containing ~5500 water molecules. The net charge of the peptides was neutralized with sodium or chloride ions. Each system was initially subject to energy minimization, followed by 1 ns of MD simulation in the NPT ensemble, with position restraints on the backbone atoms. After this initial relaxation, each system was simulated for 100ns in the NPT ensemble. The trajectories obtained from these 100ns runs were used for subsequent data analysis.

Table 4.3 The small peptides chosen for the force field validation.

ID	Sequence	Length	T
3EK	Ac-AEAAAKEAAAKEAAAKA-NH_2	16	274K
4EK	Ac-AEAAKAEAAKAEAAKA-NH_2	17	274K
3KI	Ac-AAAAKAAAAKAAAAKA-NH_2	16	274K
6KI	Ac-AKAAKAKAAKAKAAKA-NH_2	16	274K
pepIII	Ac-AETAAAKFLRAHA-$CONH_2$	13	276K
AAQAA	Ac-AAQAAAAQAAAAQAA-NHM	15	277K/300K
2JOF	DAYAQWLKDGGPSSGRPPPS	20	300K
Trp-Cage	NLYIQWLKDGGPSSGRPPPS	20	282K
GB1m1	GEWTYDDATKTATVTE	16	300K
GB1m2	GEWTYNPATGKFTVTE	16	300K
GB1m3	KKWTYNPATGKFTVQE	16	300K
GB1p	GEWTYDDATKTFTVTE	16	300K
HP5A	KKYTWNPATGKATVQE	16	300K
HP5F	KKYTWNPATGKFTVQE	16	300K
HP5W	KKYTWNPATGKWTVQE	16	300K
HP5W4	KKWTWNPATGKWTWQE	16	300K
Trpzip2	SWTWENGKWTWK-NH_2	12	288K

4.3.4 Globular Proteins

MD simulations of globular proteins, such as lysozyme, ubiquitin, and Protein G, were also performed using the Kirkwood Buff Force Field models. Simulations were initiated from PDB solvated in 80 Å cubic water boxes containing ~12500 water molecules. The net charge of the proteins was neutralized with sodium or chloride ions. Each system was initially subject to energy minimization, followed by 3 ns of MD simulation in the NPT ensemble during which the temperature was increased linearly from 100K to 300K, with applying position restraints to the backbone atoms. After this initial relaxation, each system was simulated for 100ns in the NPT

ensemble. The trajectories obtained from these 100ns runs were used for subsequent data analysis.

Table 4.4 The small globular proteins used for force field validation.

Protein	PDB ID	Source	Length	T
Barnase	1A2P	X-ray	108	287K
Barstar	1BTA	NMR	89	300K
CheY	1CYE	NMR	129	300K
FKBP12	1FKS	NMR	107	300K
Lysozyme	2LZT	X-ray	129	300K
ProteinG	3GB1	NMR	56	298K
RNaseA	2AAS	NMR	124	300K
RNaseH	2RN2	X-ray	155	300K
Ubiquitin	1D3Z	NMR	76	308K

4.3.5 MD Simulation

All condensed phase molecular dynamics simulations were performed with the SPC/E water model using the GROMACS program.[40,41] The force field parameters are presented in Tables 4.1 and 4.4. The solute geometries were constrained using Lincs, while the water geometry was constrained using SETTLE.[42] A cutoff of 1.5 nm was employed with a non-bonded pair list update of every 10 steps. Long range electrostatic interactions were evaluated using the PME approach.[43] There was no scaling of atomic charges in the condensed phase (relative permittivity of unity). The LJ and Coulomb 1-4 interactions were scaled by 0.1 and 0.5, respectively, as used for the gas phase calculations. The starting structure for dipeptide simulations was the fully extended backbone conformation. All dipeptides were solvated by SPC/E[44] water in cubic box of length 3.5 nm.

4.4 Results and Discussion

4.4.1 Backbone (KBFF-v1, KBFF-v2)

4.4.1.1 General Fitting: KBFF-v1

The QM gas phase rotational surfaces and the corresponding fitted surfaces are displayed in Figure 4.5 for the glycine, alanine and proline dipeptides. This gives rise to a version 1 of the KBFF models (KBFF-v1). The KBFF-v1 map for the glycine dipeptide was obtained with just three terms for φ and three terms for ψ dihedrals. The fitted map appears to be very reasonable with the major differences occurring in regions of very high energy. Using the φ/ψ potentials derived from the glycine dipeptide map, an additional thréeandφ threeψ´ potentials were derived by fitting to the alanine dipeptide map as shown in Table 4.5. Again, the two maps are in good agreement. The map generated for the proline dipeptide using the potentials developed for the glycine and alanine dipeptides is also reasonable. From these results it appears a scaled condensed phase charge distribution can be used to obtain reasonable fits to the QM maps using only a few torsional potentials.

The true test for the torsional potentials of peptides and proteins is the corresponding distributions obtained in the condensed phase. We performed simulations of all three dipeptides in SPC/E water using the parameters presented in Tables 4.1 and 4.5. No scaling of the charge distribution was used for the condensed phase simulations. The φ/ψ probabilities maps obtained from these simulations, together with the distributions obtained from analysis of crystal structures in the Protein Data Bank, are displayed in Figure 4.4. All the dihedral potentials are described by a simple Fourier series: $V = k [1 + \cos(n\varphi - \delta)]$.

Figure 4.5 The gas phase potential energy surfaces for glycine (top), alanine (middle) and proline (bottom) dipeptides. The QM surfaces (LMP2/cc-pVQZ//MP2/6-31G*) are shown on the left, while the KBFF surfaces are displayed on the right. The KBFF maps were generated using the solution phase charges and a relative permittivity of 1.43 to represent a scaling of the solution phase charges to mimic the gas phase charges.

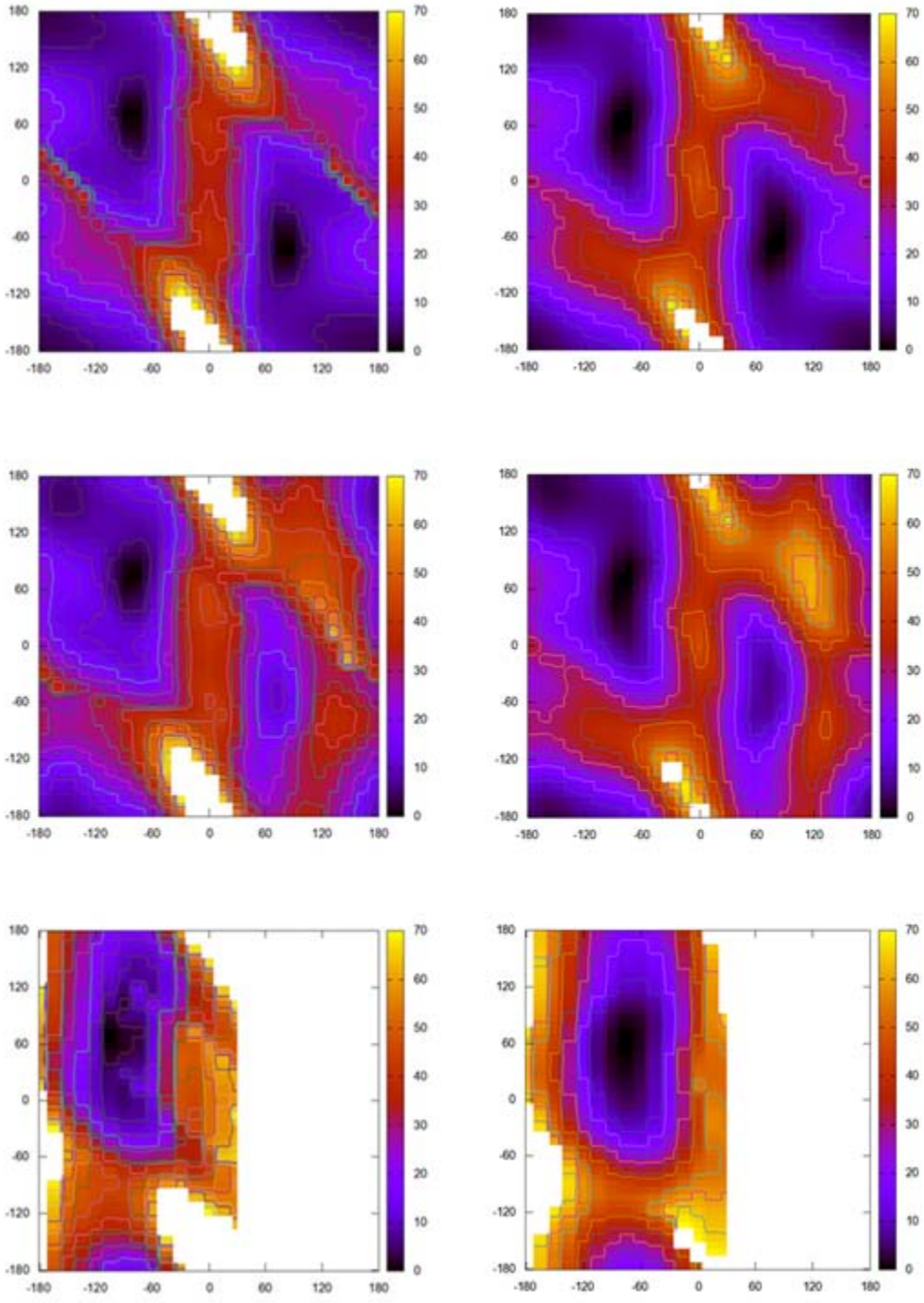
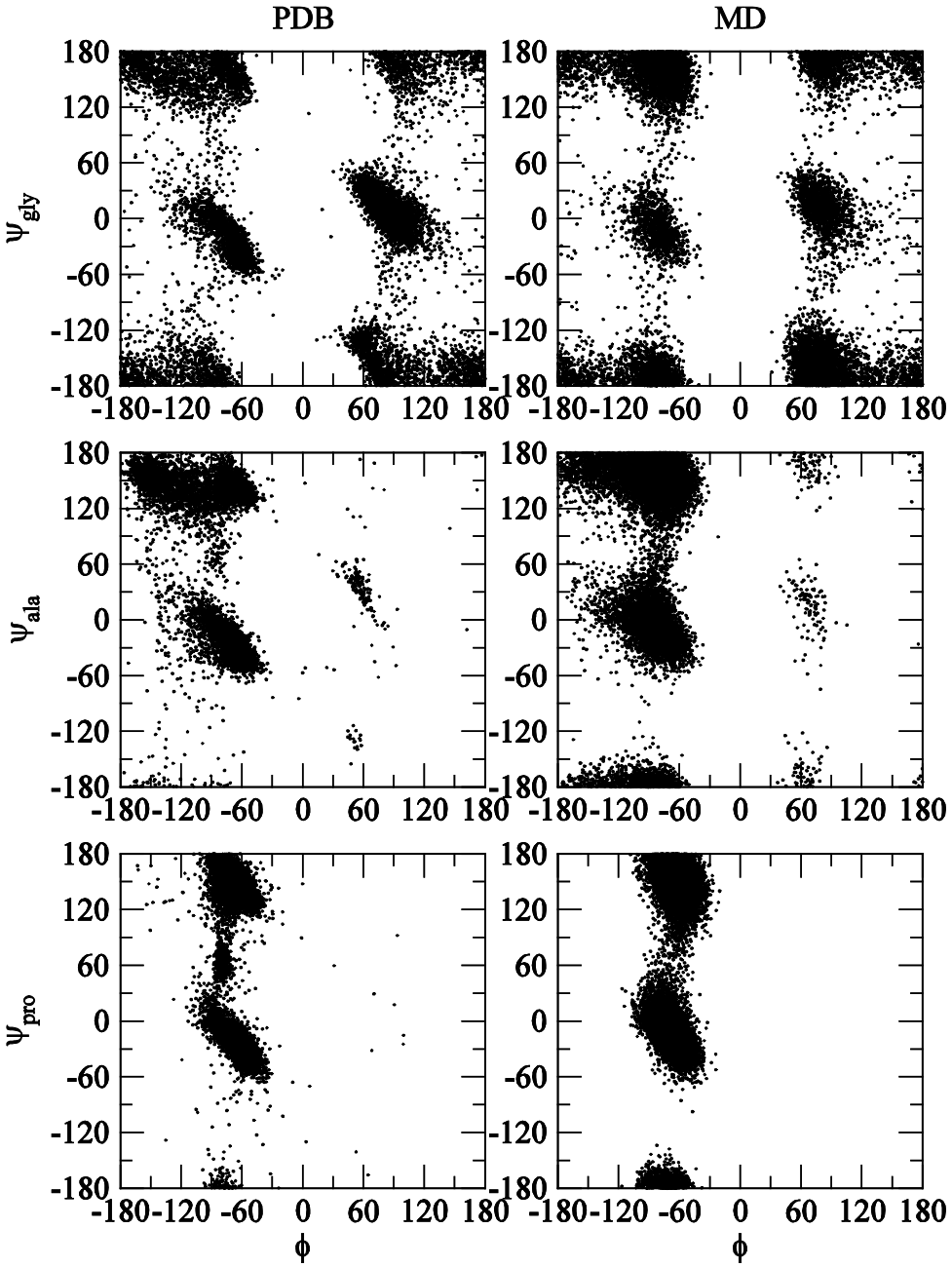


Table 4.5 Torsional parameters obtained from a fit to the gas phase maps (KBFF-v1).

Dihedral	k_ϕ (kJ/mol)	n	ϕ_s (degree)
C-N-C^α-C (φ)	-1.903	1	0.0
	1.366	2	0.0
	2.089	3	0.0
N-C-C^α-N(ψ)	5.744	1	0.0
	-4.034	2	0.0
	-0.398	3	0.0
C-N-C^α-C^β (φ´)	-2.471	1	0.0
	-1.226	2	0.0
	-0.611	3	0.0
N-C- C^α-C^β (ψ´)	2.108	1	0.0
	0.086	2	0.0
	0.332	3	0.0
C^α-C-N-C^α (ω)	7.00	1	0.0
	44.00	2	180.0
C-C-N-H	1.75	3	0.0
C-C-C/N-C/N (Pro χ1 – χ4)	7.50	3	0.0

While there is no *a priori* reason why the PDB and solvent distribution should be exactly the same, it is comforting that they sample essentially the same minima on the φ/ψ surface. The alanine distributions should be representative of the majority of amino acids. Here, we observe sampling in the major α and β regions, and also in the vicinity of C7eq which connects these two regions. This suggests facile inter conversion between the two regions. A rough analysis of a 50 ns simulation indicated approximately 0.65 ns elapsed, on average, between α→ β and β → α transitions. The major difference between the PDB and KBFF-v1 distributions is that the latter appears more extended in each region. In particular, the increased sampling of ψ between -150º and -180º is noticeable for the KBFF-v1 model. Overall, however, the distributions appear to be very reasonable and lend support to the assumption that the torsional potentials are transferable from the gas phase to the liquid phase, whereas the charge distributions are clearly not.

Figure 4.6 Ramachandran maps for glycine (top), alanine (middle) and proline (bottom) obtained from analysis of the Protein Data Bank (left), and from explicit solvent simulations of the corresponding dipeptides at 300 K using the KBFF-v1 models (right). The simulated distributions for the glycine and proline dipeptides correspond to 50 ns simulations. The simulated distribution for the alanine dipeptide corresponds to the 50 ns REMD simulation. Approximately 15000 data points are displayed in each case.



It is informative to compare the present results with the QM/MM simulation of the alanine dipeptide performed by Hu et al.[45] In that study the QM/MM simulations generated probability distributions which resembled the PDB distribution, and in contrast to the force fields available at the time. In particular, they observed a 45º slant in the distribution for the α region, and a significant population in the "pass" region located close to C7eq. Both of these features are apparent in our simulated distributions for the alanine dipeptide (Figure 4.6). Using the region definitions from Figure 4.6 of Hu et al,[45] we find the current KBFF-v1 model results in populations of 35, 4, 55, 5 and 1% for the beta, pass, alpha R, alpha L and state 4, respectively. The average populations from Table II of Hu et al. are 52, 14, 29, 3 and 1%, respectively. The KBFF-v1 model is in reasonable agreement, especially when considering that the QM/MM simulations were performed with different solute MM and solvent force fields, but with a small redistribution between the beta and pass regions. Using the same region definitions the PDB data provides populations of 65, 2, 33, 1 and 1%, respectively, which is in acceptable agreement with the KBFF-v1 derived populations.

In this study we have advocated the use of a scaled effective condensed phase charge distribution in fitting the gas phase potential energy surfaces. If one uses the full condensed phase charge distribution then the global minimum at C7eq, which involves a favorable NH to CO interaction, tends to become overly stable. Hence, during the fitting procedure one generates torsional potentials which tend to favor the α and β regions compared to those obtained with the scaled charge distribution. Consequently, when using these potentials in an aqueous environment, where the solvent destabilizes the C7eq region,[14,45] one can over stabilize both the α and β regions of the surface.

4.4.1.2 Improved ϕ, ψ Dihedrals: KBFF-v2

In general, reasonable backbone torsion parameters were obtained by fitting the QM map with a set of simple torsional potentials. However, this approach might not be sufficient to ensure a good balance between the different secondary structures in the condensed phase. This requires further simulation and possible refinement. The general distribution of Ramachandran maps looks similar to the PDB distribution, but not exactly the same. For example, the α-helix region is a little larger than observed in the PDB data, while in contrast the β-sheet region at ϕ, ψ (-180, 180) seems smaller. More importantly, initial simulations of the peptides listed in Table 4.3 indicated a bias towards the alpha helix conformation. These observations motivated the further

adjustment of the backbone dihedral parameters to better reproduce the observed experimental behavior. We label this refined force field KBFF-v2. This is considered our best force field to date.

We have attempted to improve the backbone dihedrals by using the same torsional potential functional forms and assuming that only parameters associated with the ϕ (C_{-1}-N-C_α-C) and ψ (N-C_α-C-N_{+1}) dihedrals need to be optimized. Our initial attempts involved fitting the gas phase map with the addition of an n = 1 - 6 Fourier series for both dihedrals with a total of 24 parameters (one force constant and one phase for each term), and allowing the phase shift to adopt any value. Unfortunately, the results from the peptide simulations were still not very satisfying in that a bias to the alpha helix conformation was still present. So, we decided to use a few terms and fix the phase values at -150 and 150 for ϕ, ψ torsions, respectively, and simply vary the force constants until the peptide simulations were in better agreement with the experimental data. Table 4.6 indicates the final parameters and additional terms used for ϕ, ψ dihedrals.

Table 4.6 The additional torsional terms for ϕ and ψ dihedrals leading to KBFF-v2.

	k_ϕ	n	ϕ_s
ϕ	-4.0	1	-150.0
	-1.0	6	-150.0
ψ	-1.0	1	150.0
	-1.0	2	150.0

Table 4.7 Comparison of the ϕ and ψ dihedral distributions for aqueous simulations of the alanine dipeptide obtained with different force fields.

	C5 (%)	PPII + $C7_{eq}$ (%)	α_R (%)	α_L + $C7_{ax}$ (%)
KBFF-v1	5	34	55	6
KBFF-v2	21	43	35	1
Amber99SB	21.4	52.4	21.2	5

Due to the lack of QM based results in condensed phase, we have compared many of the current force field results with Amber99SB,[15] which is believed to perform very well for proteins

simulations in solution.[14,46] Table 4.7 provides the population of conformations for the alanine dipeptide in water using our original (KBFF-v1) and improved (KBFF-v2) KBFF models compared with Amber99SB force field. The main difference between the two KBFF models is the increased population of the β distribution in the C5 region at the expense of the alpha region. Comparisons of Ramachandran map for the alanine dipeptide are shown in Figure 4.7, and the corresponding free energy surfaces are displayed in Figure 4.8. According to these results, one can also observe that the C5 region is stabilized by sacrificing the helix population. The KBFF-v2 results are closer to the Amber populations. However, it was not our intent to exactly match the population from Amber99SB. In fact, some of the peptide results (see later) obtained with the Amber force field suggests that small helices are somewhat too unstable.

Figure 4.7 Comparison of the Ramachandran maps for the alanine dipeptide obtained from explicit solvent simulations at 300 K for the Amber99SB (left) and KBFF-v2 (right) models.

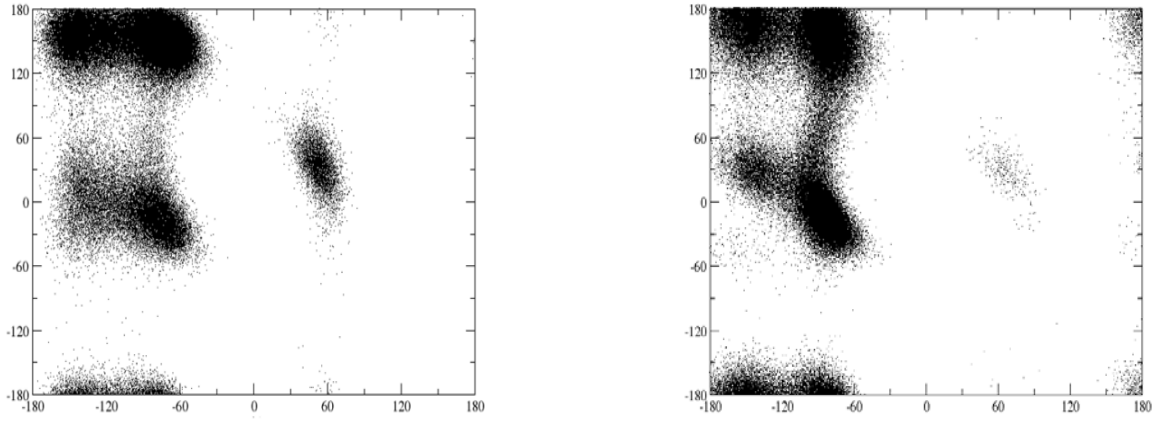
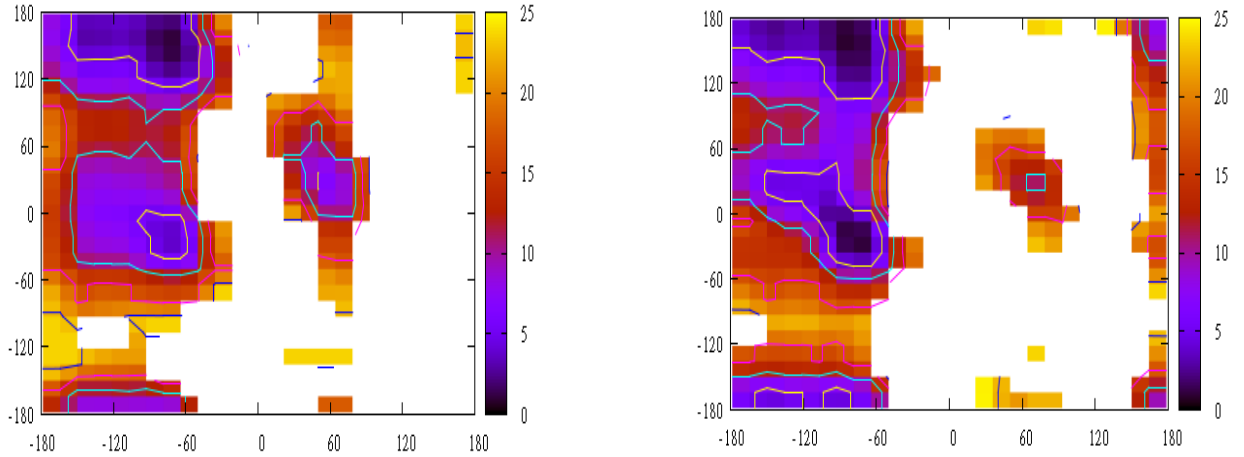


Figure 4.8 Comparison of free energy surfaces for the alanine dipeptide obtained from explicit solvent simulations at 300 K using the Amber99SB (left) and KBFF-v2 (right) models.



A recent NMR study of poly-alanine peptides in water has generated scalar coupling constant data which is sensitive to rotation around φ and ψ.[39] It has also been shown that achieving good agreement with the experimental data is nontrivial for existing force fields.[5,20,47] We have therefore performed a series of simulations of the Ala_3, Ala_5, Ala_7, Gly_3 and Val_3 peptides in water at 300 K and computed the scalar coupling constants according to the Karplus equation used in the original NMR study.[48-51] Only couplings for the central residues were investigated to help eliminate specific terminal effects. The simulated and experimental coupling constants are displayed in the Figures 4.9 - 4.13. In addition, a series of coupling constants from Ala_3 as a function of temperature were displayed in Figure 4.14. The results suggest that the KBFF-v2 model provides an excellent description of the distribution around φ and ψ for most cases, with only a few exceptions. For the case of polypeptide Ala_5, the overall χ^2 was 1.7 for KBFF-v1 and 2.5 for KBFF-v2. Generally speaking, the overall χ^2 value of KBFF-v1 is lower than any of the previously tested force fields[14] as shown in Table 4.8, while the overall χ^2 value of KBFF-v2 is a little larger than previous version but still acceptable. It is possible that the agreement with experiment for the couplings probing rotation around φ and ψ could be improved

with a small adjustment to the φ/ψ distribution. We have not pursued this here as we believe it is unwise to focus solely on a single set of experimental results when making such adjustments.

Table 4.8 Comparison of χ^2 from polypeptide Ala_5 among a series of different force fields.[14]

Force Field	χ^2 (Ala_5)
Amber03	1.6
Amber99SB	4.2
AmberGS	1.9
Charmm27/cmap	2.2
OPLS-AA/L	2.0
Gromos53a6	2.3
Gromos43a1	1.6
KBFF – v1	1.7
KBFF – v2	2.5

Calculation is based on Karplus equation[52] using $\chi^2 = N^{-1}\sum_{j=1}^{N}\left(\langle J_j\rangle_{sim} - J_{j.expt}\right)^2/\sigma_j^2$.

Figure 4.9 Simulated (KBFF-v2) and experimental scalar coupling constants for Ala_3 in water at 300 K. Coupling constants sensitive to rotation around φ (black) and rotation around ψ (red) are indicated here.

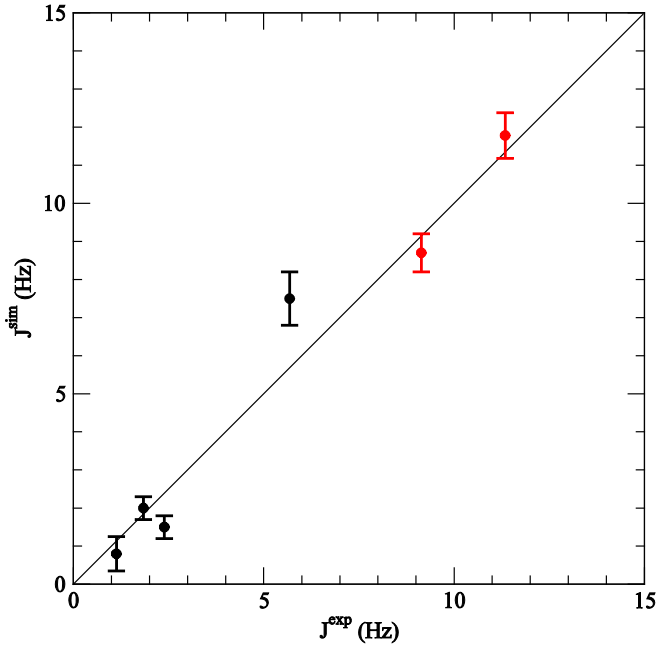


Figure 4.10 Simulated (KBFF-v2) and experimental scalar coupling constants for Ala_5 in water at 300 K. Coupling constants sensitive to rotation around φ (black) and rotation around ψ (red) are indicated here.

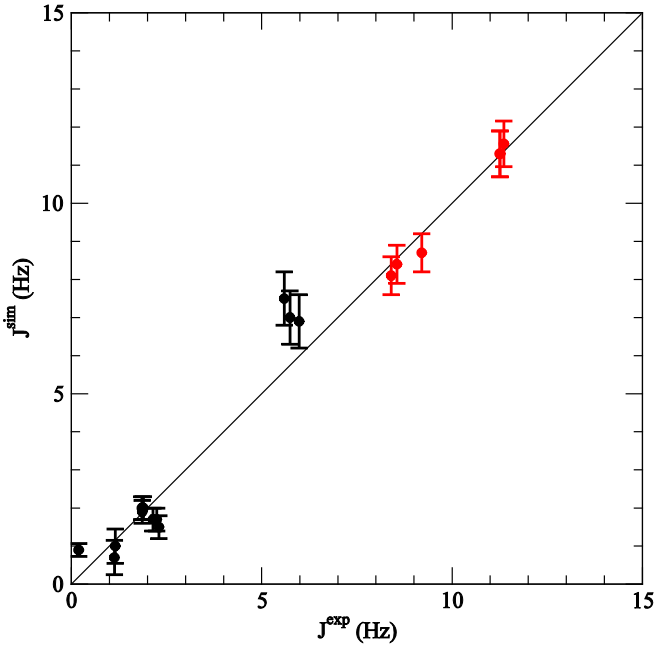


Figure 4.11 Simulated (KBFF-v2) and experimental scalar coupling constants for Ala_7 in water at 300 K. Coupling constants sensitive to rotation around φ (black) and rotation around ψ (red) are indicated here.

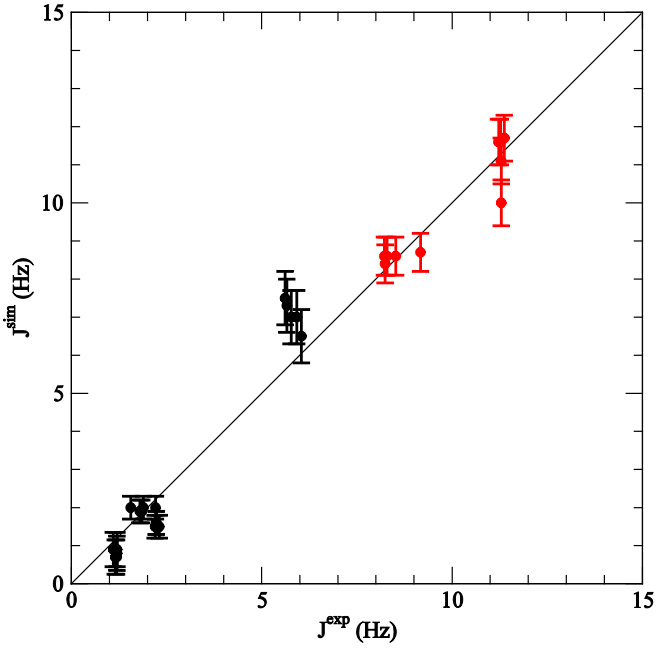


Figure 4.12 Simulated (KBFF-v2) and experimental scalar coupling constants for Gly_3 in water at 300 K. Coupling constants sensitive to rotation around φ (black) and rotation around ψ (red) are indicated here.

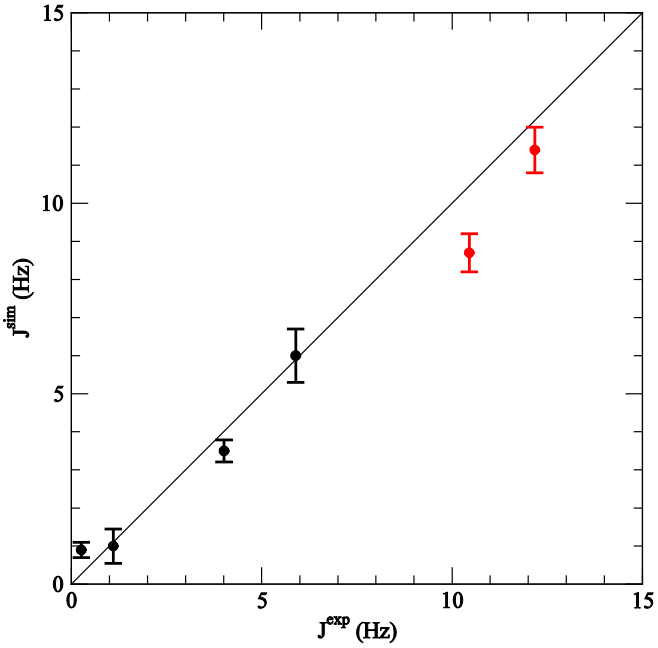


Figure 4.13 Simulated (KBFF-v2) and experimental scalar coupling constants for Val_3 in water at 300 K. Coupling constants sensitive to rotation around φ (black) and rotation around ψ (red) are indicated here.

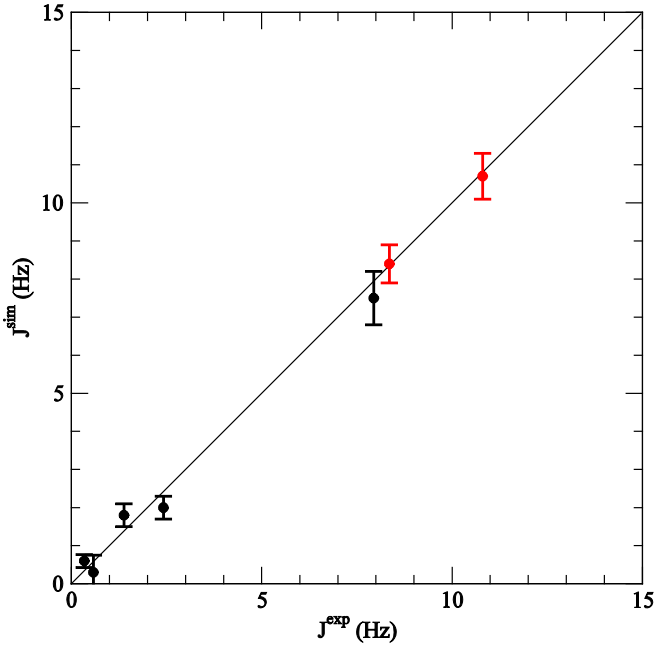
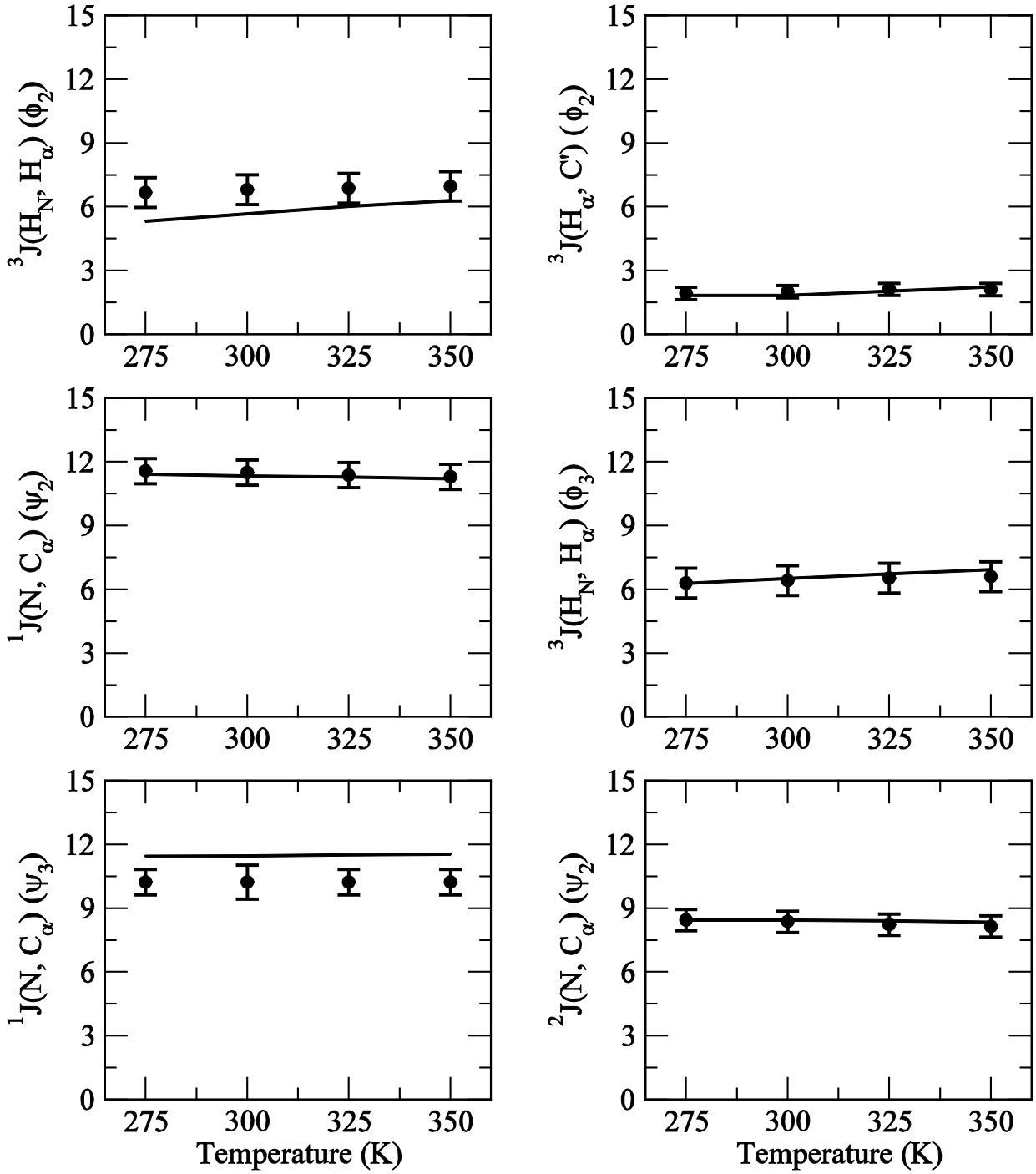


Figure 4.14 Simulated (KBFF-v2) and experimental J coupling constants for Ala_3 in water at 275 to 350 K. The black lines correspond to coupling constants obtained from the experimental data,[39] and the points are from a simulation using the KBFF model.



4.4.2 Sidechain (KBFF-v1)

4.4.2.1 General Fitting

In order to develop and improve the performance of the KBFF models for biological systems we have also examined the torsion potentials for the amino acid side chains. Again, we use a simple Fourier series, $V = k_\phi [1 + \cos (n\varphi - \phi_s)]$, to describe the potential energy as a function of rotation. The initial force constants for the torsional potentials were determined by fitting the results of literature ab initio 6-31G* calculations on small molecules.[37] The results are displayed in Table 4.9 and display excellent agreement - the difference between two data sets being less than 0.6 kJ/mol. We believe, and it is commonly assumed, that these parameters are transferrable to related systems and can therefore be used to describe the rotational energies of side chains in common amino acids and peptides.

Table 4.9 The torsional parameters (KBFF-v1) for model compounds representative of amino acid sidechains. The relative energies (kJ/mol) for different conformations (and barriers) are also displayed and compared with QM values.[35]

Dihedrals	k_ϕ	n	ϕ_s	conf	KB	6-31G*
	0.25	1	0.0	0	26.0	25.9
C-C-C-C	-0.25	2	0.0	60	4.3	4.2
butane	7.25	3	0.0	120	15.3	15.2
	-	-	-	180	0.0	0.0
	-0.75	1	0.0	0	22.6	22.6
C-C-C-O	0.50	2	0.0	60	0.0	0.00
propanol	8.50	3	0.0	120	16.4	16.4
	-	-	-	180	0.3	0.3
	0.75	1	0.0	0	7.5	7.5
C-C-O-H	0.50	2	0.0	60	0.5	0.5
ethanol	3.00	3	0.0	120	5.5	5.5
	-	-	-	180	0.0	0.0
C-C-S-H	0.00	1	0.0	0	30.9	30.9
ethanethiol	0.75	2	0.0	60	0.0	0.0

	3.25	3	0.0	120	6.4	6.4
	-	-	-	180	1.3	1.3
C-C-S-C $CH_3CH_2CH_2SCH_3$	2.00	1	0.0	0	18.9	18.8
	1.25	2	0.0	60	2.4	2.3
	4.25	3	0.0	120	7.7	7.6
	-	-	-	180	0.0	0.0
C-S-S-C CH_3SSCH_3	7.50	1	0.0	0	46.1	46.1
	16.00	2	0.0	90	0.0	0.0
	-	-	-	180	22.4	22.4
C-C-Car-Car ethylbenzene	0.75	2	0.0	0	6.0	5.9
	-	-	-	90	0.0	0.0
C-C-Car-Car 3-ethylindole	0.25	1	0.0	0	0.0	0.0
	-1.25	2	0.0	60	4.9	4.9
	-2.75	3	0.0	120	0.9	0.9
	-	-	-	180	17.8	17.8
C-C-C-N 5-ethylimidazole	2.00	1	0.0	0	8.3	8.1
	0.25	2	0.0	60	0.0	0.0
	3.00	3	0.0	120	4.8	4.6
	-	-	-	180	0.6	0.6
C-C-C-N propanamide	2.75	1	0.0	0	7.6	7.1
	-4.75	2	0.0	180	0.0	0.0
C-C-C-O propanoate ion	-4.00	2	0.0	0	0.0	0.0
	-	-	-	90	3.0	3.0
C-C-N-C ethylguanidinium ion	-0.50	1	0.0	0	27.9	28.0
	-0.25	2	0.0	60	4.5	3.9
	3.25	3	0.0	120	6.4	6.9
	-	-	-	180	0.0	0.0
O-C-O-H acetic acid	2.50	1	0.0	0	24.5	24.5
	-20.00	2	0.0	90	52.4	52.5
	-	-	-	180	0.0	0.0

4.4.2.2 KBFF-v1 χ1 Dihedrals

One of the more important side chain dihedrals which deserves additional attention is the χ1 dihedral defined by N-C^{α}-C^{β}-C^{γ} atoms. The initial parameters for the χ1 dihedrals of each amino acid were developed using a similar approach to that in Section 4.4.2.1. The dihedral around C^{α} and C^{β} can be considered to include two terms χ^1 (N-C^{α}-C^{β}-C^{γ}) and $\chi^{1'}$ (C-C^{α}-C^{β}-C^{γ}). Since these two terms look similar to the dihedrals found in molecules with the C-C-C-N (propylamine) and C-C-C-C (butane) linkages, parameters for the two χ1 and $\chi^{1'}$ dihedrals were taken from QM calculations on these molecules.[35] However, since both χ1 and $\chi^{1'}$ contribute to the rotational energy surface around C^{α} and C^{β}, we decided to take half the value of each original dihedral potential term to avoid overestimating the dihedral energy in this region. The final parameters are shown in Table 4.10.

Table 4.10 General parameters (KBFF-v1) for χ1 (N-Cα-Cβ-Cγ) and χ1′ (C-Cα-Cβ-Cγ) dihedrals.

Chi1	k_ϕ	n	ϕ_s
C-C^{α}-C^{β}-C^{γ}	0.125	1	0.0
	-0.125	2	0.0
	3.625	3	0.0
N-C^{α}-C^{β}-C^{γ}	-0.375	1	0.0
	0.250	2	0.0
	4.250	3	0.0

Unfortunately, this (admittedly simplistic) approach did not provide reasonable agreement with conformational populations observed in the PDB database. Here, we found that the χ_1 torsional angle is sensitive to each amino acid residue type and therefore the χ_1 distributions for many residues were significantly different from the data from the rotamer library.[36] Hence, we attempted to improve the agreement by including additional torsional parameters. Our initial attempts at a reduced set of torsional parameters involved dividing the side chain χ1 dihedrals into four types based on residue character:

1) Hydrophobic side chain, e.g. Val, Leu, Ile, Met

2) Hydrophilic side chain with carbonyl group, e.g. Asn, Asp, Arg, Glu

3) Hydrophilic side chain with hydroxyl or amine group Thr, Cys, Lys

4) Side chain with carbon ring Phe His Trp Tyr

and then refining the χ1 dihedrals by residue type. However, this approach still provides relatively poor agreement with the experimental rotamer populations. Hence, we were forced to improve the χ1 distributions by individually modifying the χ1 dihedral parameters for each amino acid. It is believed that the distribution of structures in the PDB shows a good approximation for that observation from MD simulation.[16,53-55] Therefore, we compared the results from MD simulations to the corresponding populations in the PDB from rotamer library to improve the force field results. More specifically, we performed a series of MD simulations of short peptides with sequences ACE-Ala-X-Ala-NHM, where X is any natural amino acid apart except for Gly, Ala and Pro. From these simulations we calculated the relative populations of the plus (+60), minus (-60) and trans (180) χ_1 rotamers for each residue and compared them to the relative populations observed for the same residue in PDB.[36] We decided to adopt an REMD approach in the simulations to help overcome any possible large energy barriers by enhanced sampling In addition, as some amino acid like Asp and Asn display complicated rotameric preferences for χ_2, and their χ_1 and χ_2 torsions appeared to be strongly coupled, we examined the full energy profile for both χ_1 and χ_2 in some cases. According to Table 4.11 providing the final additional χ1 torsional parameters added to each amino acid, Table 4.12 provides the simulated and experimental conformational populations for the KBFF-v1 models.

Table 4.11 The additional torsional parameters (KBFF-v1) used for the χ1 dihedrals (N-C^{α}-C^{β}-C^{γ}) of each amino acid.

Chi1	k_ϕ	n	ϕ_s
Arg	-1.87	1	60.0
	-3.13	2	0.0
Asn	-5.80	1	0.00
	21.27	2	60.00
	5.00	3	0.0
Asp	-1.66	1	0.00
	15.76	2	60.00
Cys	13.00	1	0.00

	-3.35	2	60.00
Gln	-0.82	1	60.00
	-3.12	2	0.00
Glu	1.00	1	60.00
	-3.40	2	0.00
His	-3.67	1	60.00
	-9.13	2	0.00
Ile	2.00	1	60.00
	1.70	2	0.00
Leu	-1.47	1	60.00
	-2.67	2	0.00
Lys	-0.87	1	60.00
	-3.00	2	0.00
Met	4.20	1	60.00
	4.20	2	0.00
Phe	2.39	1	60.00
	-4.95	2	0.00
Ser	23.90	1	0.00
	-7.87	2	60.00
	5.50	3	0.00
Thr	14.60	1	0.00
	-8.50	2	60.00
	5.00	3	0.00
Trp	7.51	1	0.00
	-0.15	2	60.00
Tyr	1.47	1	60.00
	-5.33	2	0.00
Val	0.65	1	0.00
	-2.00	2	60.00
	-7.25	3	0.00

Table 4.12 A Comparison of the χ1, χ2 distributions from REMD simulations of the blocked A-X-A peptides using KBFF-v1 with the PDB rotamer library.[36]

AA	χ1	exp (%)	sim (%)	χ2	exp (%)	sim (%)
Arg	60	11	16	60	6	3
	180	33	37	180	66	92
	-60	56	47	-60	8	5
Asn	60	17	25	0, 180	28	54
	180	29	30	-180, 0	66	46
	-60	54	30	-	-	-
Asp	60	20	15	0, 180	36	46
	180	28	25	-180, 0	64	54
	-60	52	35	-	-	-
Cys	60	23	20	-	-	-
	180	26	30	-	-	-
	-60	50	50	-	-	-
Gln	60	6	6	60	14	33
	180	27	24	180	55	37
	-60	56	70	-60	20	30
Glu	60	8	7	60	12	33
	180	34	43	180	62	59
	-60	58	50	-60	16	8
His	60	14	20	60, 180	40	41
	180	34	30	-180, -60	54	59
	-60	52	50	-	-	-
Ile	60	14	15	60	4	4
	180	10	8	180	81	90
	-60	76	77	-60	15	6
Leu	60	1	2	60	32	36
	180	33	33	180	61	64
	-60	66	65	-60	0	0

	60	6	8	60	5	2
Lys	180	38	39	180	68	96
	-60	56	53	-60	9	2
	60	5	6	60	7	6
Met	180	22	34	180	56	88
	-60	60	60	-60	24	6
	60	13	18	20, 150	46	48
Phe	180	33	26	-100, -20	38	52
	-60	54	56	-	-	-
	60	49	40	-	-	-
Ser	180	22	20	-	-	-
	-60	29	40	-	-	-
	60	49	45	-	-	-
Thr	180	7	15	-	-	-
	-60	43	40	-	-	-
	60	18	18	-50, 50	8	10
Trp	180	35	30	50, 180	56	40
	-60	47	52	-120, -50	33	50
	60	13	20	50, 150	47	52
Tyr	180	35	30	-150, -50	52	48
	-60	52	50	-	-	-
	60	6	8	-	-	-
Val	180	73	69	-	-	-
	-60	20	23	-	-	-

In nearly all cases the KBFF-v1 parameters for side chains could reproduce the experimentally observed population distributions.

4.4.3 Test Cases

4.4.3.1 Small Peptides

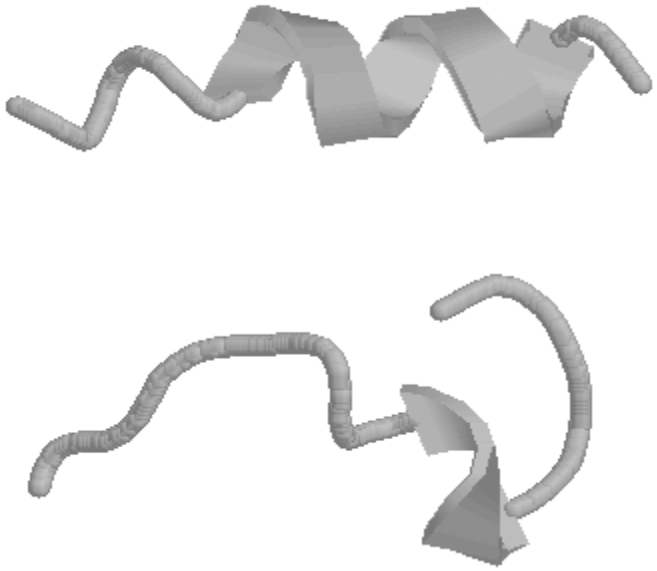
The parameter development described above led to a general force field for peptides and proteins we have termed KBFF-v1. During our initial simulations of the set of small peptides with varying degrees of alpha and beta structure, we observed a bias towards the alpha structure in many systems. This prompted a further refinement of the force field which focused on just the backbone torsional potentials (see Section 4.4.1.2) and provided our latest versions of the force field termed KBFF-v2. The following sections provide a description of the results for the test set of peptides and proteins using this latest version. To test the secondary structure preferences of the KBFF-v2 models we have simulated a series of peptides which are known to display substantial helices and sheet content as shown in Table 4.13. The helix/sheet percentage was calculated by the ratio between actual hydrogen bond frequency to ideal hydrogen bond frequency from peptide backbones.

All the helical peptides appeared to perform very well. The 3EK, 4EK 3KI, 6KI and pepIII all started from a fully helical structure, and each peptide was simulated for 100 ns and then their secondary structure was analyzed and compared with experimental data.[56]

Figure 4.15 Ten snapshots of 4EK (left) and pepIII (right) extracted for each peptides every 10 ns for a total 100ns run and compared with initial PDB structures (blue).

In the case of $(AAQAA)_3$, the experimental data[57] for this peptide suggests that the N-terminal residues are approximately 50% helical, while and the C-terminal region is disordered at 277 K. On increasing the temperature to 300 K the peptide displays essentially no helix content. Snapshots from the simulations are displayed in Figure 4.16. At low temperature the peptide maintained helical structure in the N-terminal section, while the C-terminal residues unfolded to some degree. At the higher temperature the peptide completely unfolds with only small transient regions of helix character appearing at a variety of positions in the sequence. These results are in very good qualitative agreement with the experimental data at both temperatures.

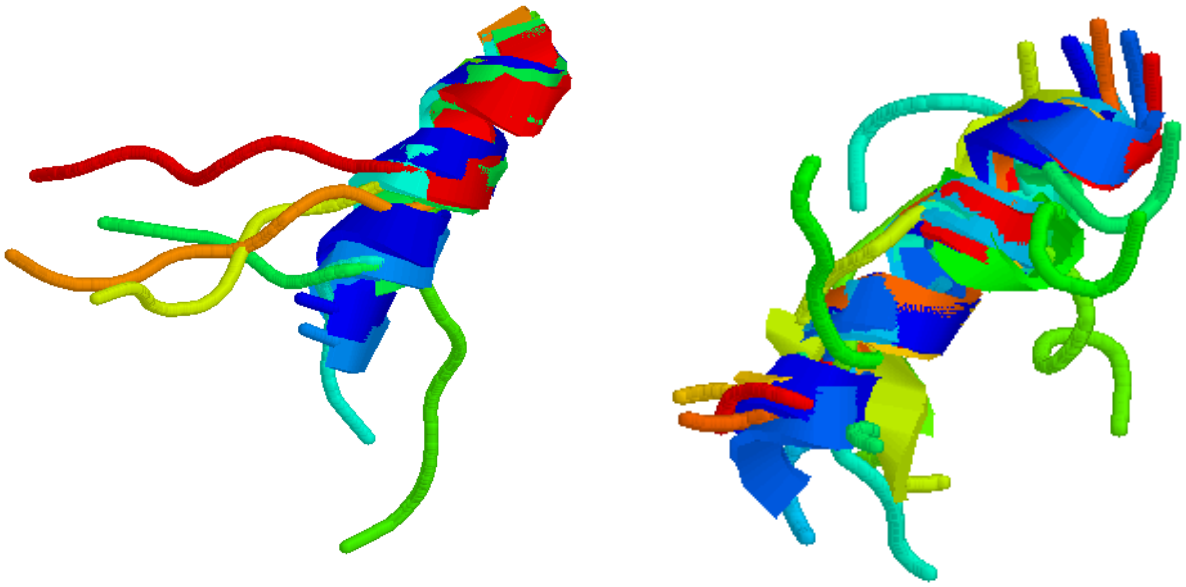
Figure 4.16 Snapshots from the KBFF-v2 simulation of $(AAQAA)_3$ at two different temperatures.



As shown in Figure 4.17 and Table 4.13, the computed helicity of 61% from KBFF-v2 is in reasonable agreement with experimental value of 50%. This compares with the helicity of 40% observed from a simulation using the Amber99SB force field. What is not clear from Figure 4.17 is that the $(AAQAA)_3$ peptide during the KBFF-v2 simulation lost most of its helicity at around 50 ns (green), and then refolded to a helix at the N-terminal keeping the C-terminal

region unfolded. In contrast, the Amber99SB simulation retains all helical structure during the first 50 ns and then during the second 50 ns more than half the length of the peptide from the C-terminus remained random coil all the time. Hence, the calculated percentages for the Amber simulation could represent a somewhat high estimate, due to the absence of reversible folding.

Figure 4.17 Comparison of $(AAQAA)_3$ at two force fields, Amber99SB (left) and KBFF-v2 (right) at 277K.



The Trp-Cage peptide contains a short α-helix together with a rigid poly-proline C-terminus and a hydrophobic Trp pocket. The simulation results for the KBFF-v2 model are shown in Figure 4.18 and compared with Amber99SB force field and also the initial PDB structure.[58] Both force fields can reproduce the half helix half coil conformation with high similarity, although the KBFF-v2 simulation appears to better capture the structure of the turn region correctly.

Figure 4.18 Ten snapshots of Trp-Cage from Amber99SB (left) and KBFF-v2 (right) extracted for each peptides every 10 ns for a total 100ns run and compared with initial PDB structures (blue).

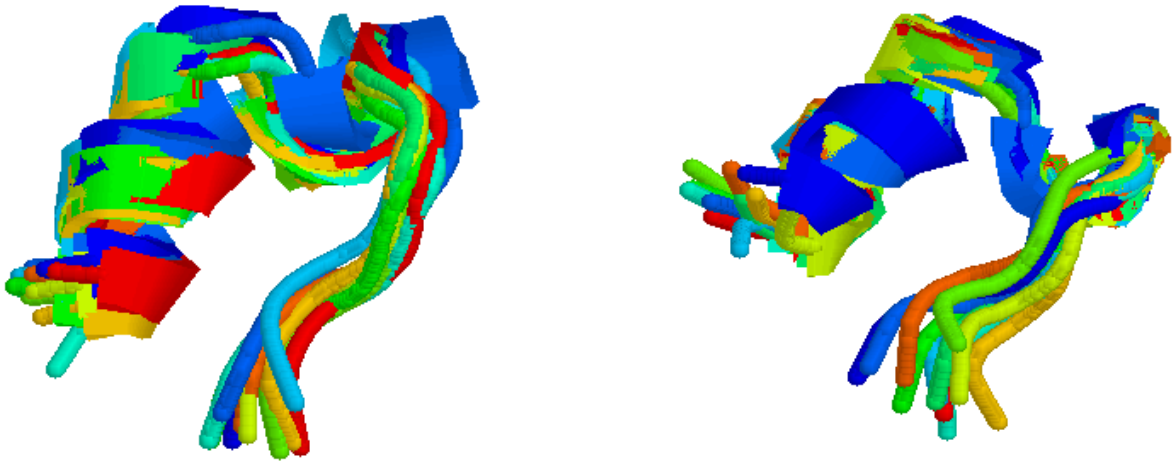


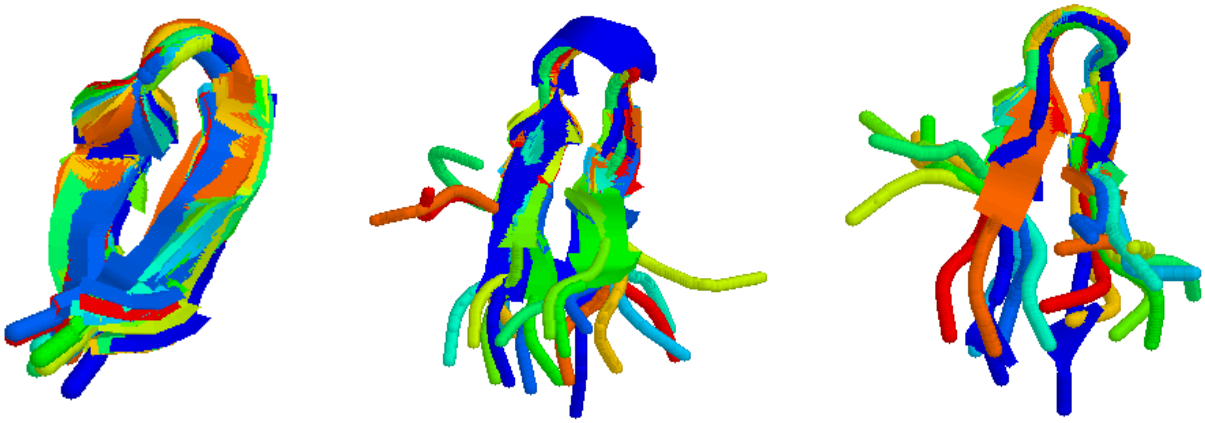
Table 4.13 KBFF-v2 simulation results for small peptides compared with experimental data.

ID	exp[53-57]	MD
3EK	20% - 50% α	70% α
4EK	80% α	83% α
3KI	80% α	72% α
6KI	25% - 50% α	64% α
pepIII	50% α	65% α
AAQAA	50% α	61% α
GB1m1	6±6% β	30% β
GB1m2	74 ± 5% β	24% β
GB1m3	86 ± 3% β	90% β
GB1p	30% – 80% β	50% β
HP5A	21 ± 10% β	31% β
HP5F	82 ± 4% β	81% β

HP5W	92 ± 2% β	77% β
HP5W4	> 96% β	82% β
Trpzip2	90% β	94% β

A series of mutants of the second hairpin from the GB1p protein have been designed and studied in detail experimentally and are observed to possess a variety of different fold stabilities. The major mutations involve replacing the native sequence DDATKT with NPATGK, and changing the N-terminal residue to Lys to stabilize the hairpin fold as indicated in Table 4.3. Other minor modifications focused on mutating several hydrophobic and aromatic residues (Val, Trp, Tyr, Phe) in order to further investigate the stability. The simulation results were compared with the experimental data.[59] The simulated stability varies from unfolded to most folded (GB1m1 < HP5A < GB1p < GB1m3), showing reasonable agreement with experiment with the exception of GB1m2. The simulation of the Trpzip2 peptide indicate that this peptide is quite stable under our KB force field, in agreement with the others studies,[60] and suggest a reasonable description of hairpin stability and aromatic residue interactions are provided by the models.

Figure 4.19 Ten snapshots of Trpzip2 (left) and GB1p (middle) and HP5A (right) from the KBFF-v2 models extracted for each peptide every 10 ns for a total 100ns run and compared with initial PDB structures (blue).



The observed hairpin percentage for GB1p is 30% to 80% based on experimental data, which is a rather large range and therefore difficult to judge the accuracy of simulation results. We compared the percentage of hairpin as shown in Figure 4.19 in order to follow the correct folding trend, and then compared with Amber force field as shown in Figure 4.20. It appears the result from the Amber99SB force field describes a reasonably rigid hairpin with over 80% population, while our results indicate much more flexibility with only 50% hairpin population. Reversible folding of the two tails was observed during the KBFF-v2 simulations.

Figure 4.20 Ten snapshots of GB1p from Amber99SB (left) and KBFF-v2 (right) extracted from each peptides every 10 ns for a total 100ns run and compared with initial PDB structures (blue).

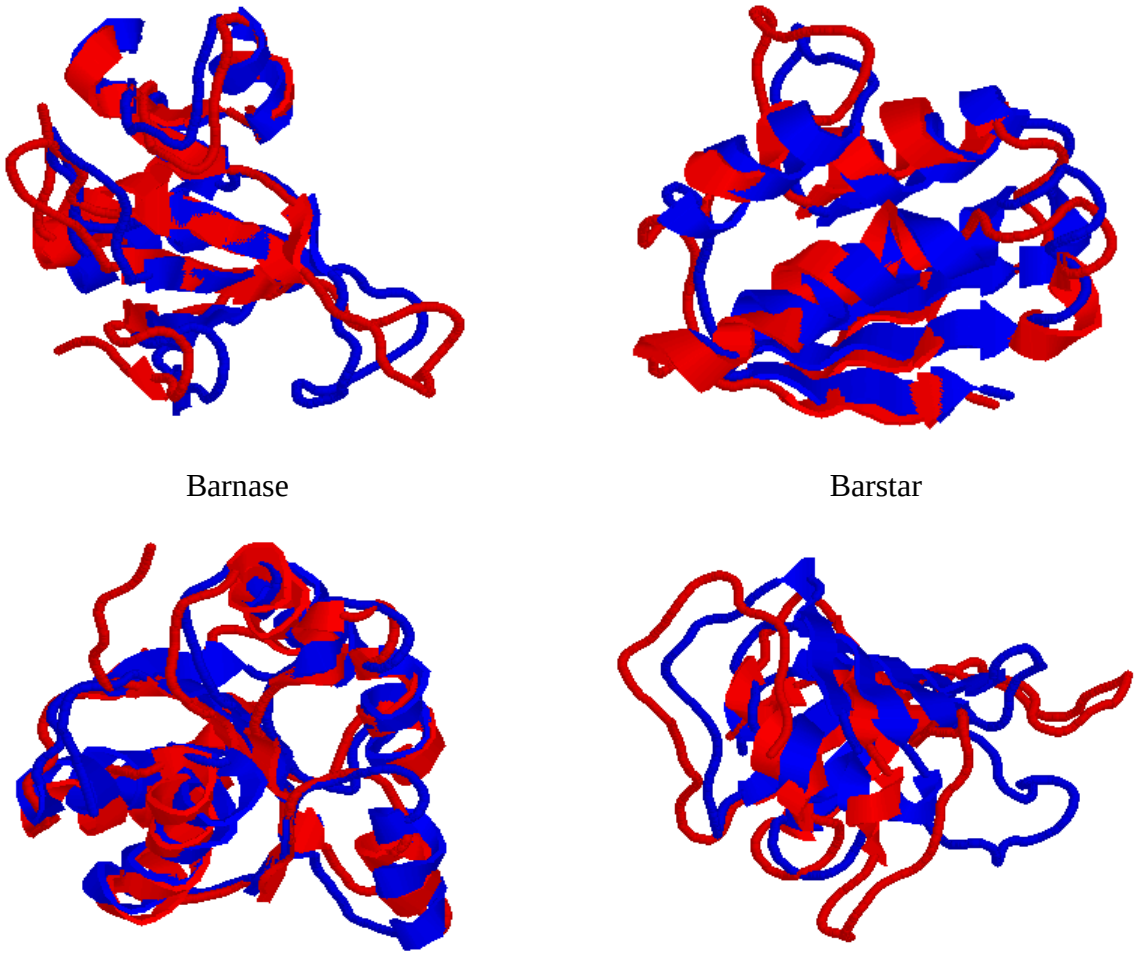
In general, it appears the KBFF-v2 models provide a reasonable balance between alpha and beta propensity in the above peptides.

4.4.3.2 Globular Proteins

In order to further probe whether the torsion potentials developed here can well reproduce the experimental results from NMR or crystal structure data, we performed relatively long-time (100ns) simulation of several selected proteins with various degrees of alpha and beta

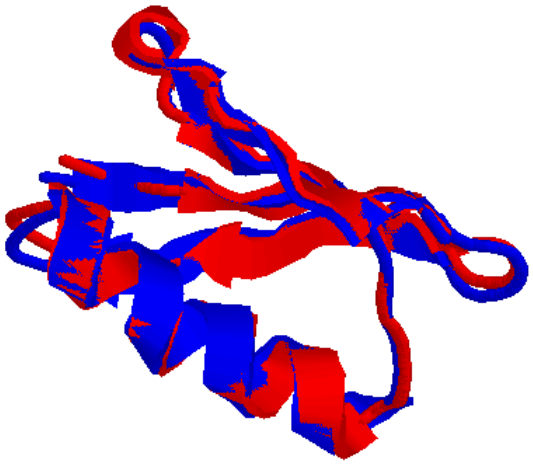
content.[20] The conformations sampled during the simulations were then compared to experimental PDB structures by calculating the average and final root-mean-square-deviation (rmsd) values, and also the rmsd of the average structure obtained over the entire trajectory. The results are presented in Table 4.10. From our simulations, we observe that the rmsd for the average structure is generally lower than the averaged rmsd values. Moreover, it is believed that the rmsd of an average structure is most appropriate for comparison with experimental data than the other two types of rmsd values.[20] It appears that over the 100 ns simulations our KB force field displayed relatively small deviations from the experimental structures, which indicates a high level of accuracy for the ϕ/ψ distributions and the nonbonded interactions in the current force field.

Figure 4.21 Initial (blue) and final (red) structures compared for each protein.

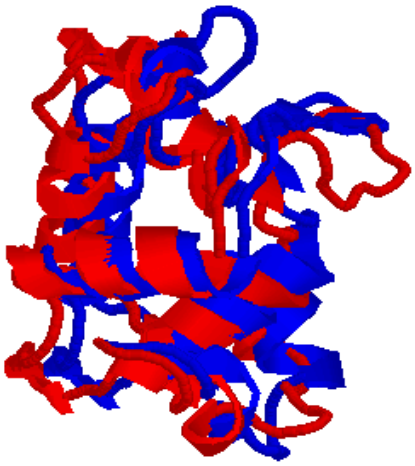


Barnase Barstar

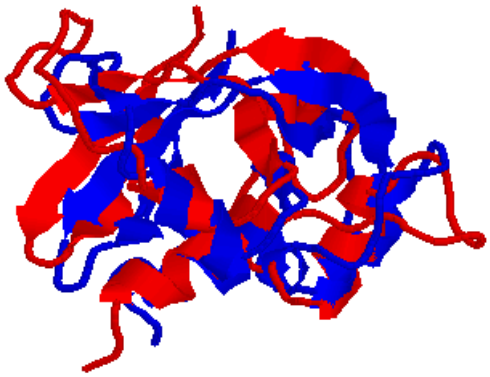
CheY FKBP12



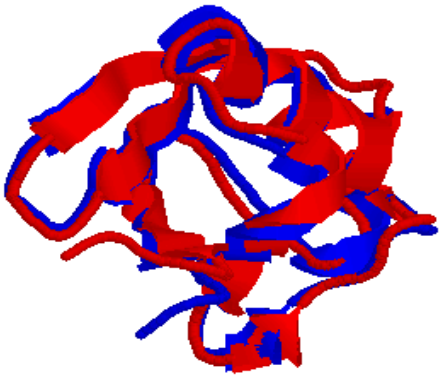
ProteinG



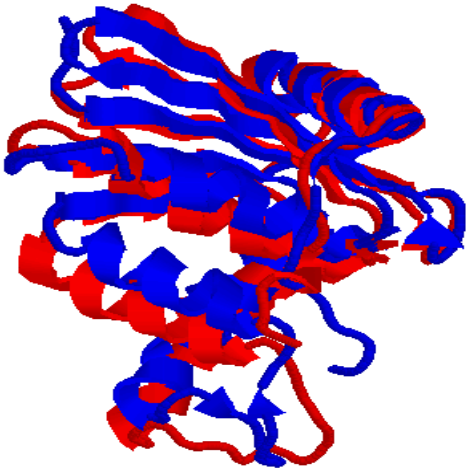
Lysozyme



RNaseA



Ubiquitin



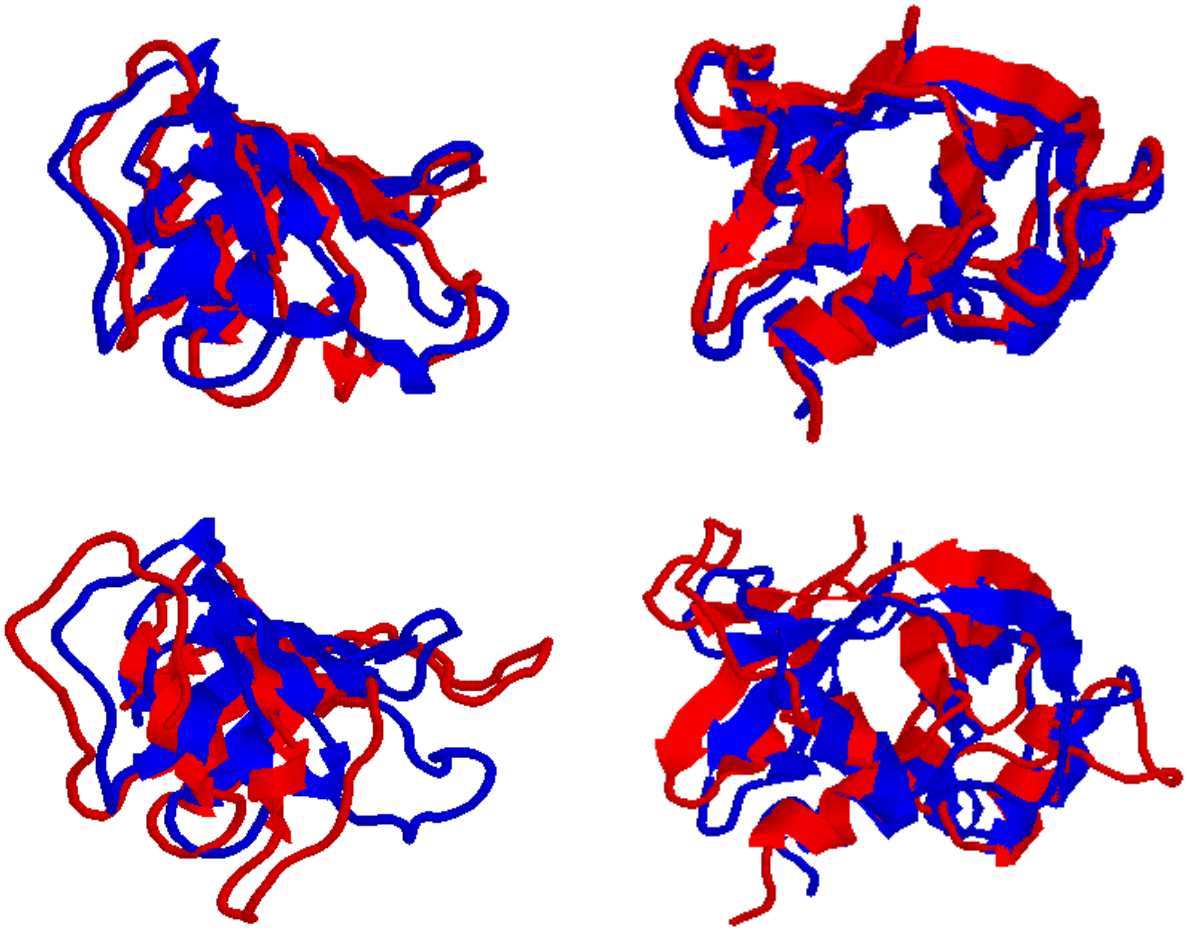
RNaseH

Table 4.14 The RMSD from the KBFF-v1 and KBFF-v2 simulations of small globular proteins.

Protein	avg. Cα rmsd (Å)		Cα rmsd of final strucutre (Å)		Cα rmsd of avg. structure (Å)	
	KB – v1	KB – v2	KB – v1	KB – v2	KB – v1	KB – v2
Barnase	3.2	2.1	3.7	3.0	2.8	1.8
Barstar	1.6	1.9	1.8	2.2	1.3	1.7
CheY	2.3	1.7	2.3	3.0	1.8	1.3
FKBP12	3.9	2.8	5.3	3.7	3.3	2.3
Lysozyme	2.6	2.5	2.9	2.8	2.3	2.1
ProteinG	1.3	0.7	1.1	0.4	1.0	0.5
RNaseA	3.5	3.0	4.8	2.8	3.0	2.7
RNaseH	2.0	2.3	2.2	2.8	1.9	1.9
Ubiquitin	1.8	1.4	1.2	1.6	1.4	0.5

Most of the proteins simulated here performed very well with RMSD values generally smaller than 3 Å, especially for the KBFF-v2 simulations. The reduction in RMSD on going from the KBFF-v1 to the KBFF-v2 was most for the FKBP12 and RNaseA proteins, indicating the important contribution of modifications in KBFF-v2. These two proteins are generally difficult to simulate accurately.[20] In Figure 4.22 we compare our results with Amber99SB simulations for these problematic proteins. Both can precisely reproduce the α-helix and β-sheet regions, the major differences appear in the loop regions where the Amber simulations perform much better than us. Thus, future efforts might have to be focused on the improvement of loop regions.

Figure 4.22 A comparison of the initial (blue) and final (red) structures of FKBP12 and RNaseA obtained from the Amber99SB (top) and KBFF-v2 (bottom) force fields.



FKBP12 RNaseA

4.5 Conclusions

Torsional potentials for the glycine, alanine and proline dipeptides that are consistent with the KBFF models for amides have been determined by fitting to QM gas phase potential energy surfaces. In doing so it was assumed that the torsional parameters derived for the gas phase are transferable to the condensed phase, while the gas phase partial atomic charges can be approximated by a scaling of the effective condensed phase charges. Using this approach one obtains very reasonable fits to the gas phase surfaces of all three dipeptides using a relatively small number of torsional terms, some of which are transferable between amino acids.

Simulation of the KBFF models in explicit solvent generated population distributions which closely match that of similar residues in the Protein Data Bank. Furthermore, a simulation of Ala_5 in explicit solvent resulted in coupling constants in good agreement with the experimental NMR data, especially for couplings probing rotation around ψ dihedrals. The fitted maps and simulation data generally display superior agreement with the available experimental data and suggest a simple approach for the improvement of force fields for biological simulations.

In addition, torsional dihedral potentials for the amino acid side chains were developed and further refined to ensure the distribution of χ1 and χ2 populations for each amino acid reasonably reproduced data from rotamer libraries. A series of peptides and proteins were then simulated to test and validate the KB force field. The final optimized KB force field, KBFF-v2, can correctly reproduce the conformational equilibria of both α-helical peptides and β-hairpin peptides. Most importantly, the simulations demonstrate that the helices/sheets generated using the current torsional potential fitting are neither too stable nor too unstable. Long time simulations of small globular proteins demonstrated the KB force field is capable of reproducing stable protein folds in general. However, more detailed structure analysis should be performed in the near future and slight adjustments might be needed for further improvement.

In summary, we have proposed a set of backbone and sidechain torsion parameters for use with the KBFF models of amino acid analogues, with further refinement to the ϕ/ψ and χ1 dihedrals. Tests on small peptides and proteins suggest the force field provides a reasonable description of the conformational preferences of small peptides and the equilibrium structure of larger proteins.

4.6 References

(1) Brooks, C. L., III; Karplus, M.; Pettitt, B. M. *Proteins: A Theoretical Perspective of Dynamics, Structures, and Thermodynamics*; John Wiley & Sons: New York, **1988**.

(2) Becker, O. M.; MacKerell, J., A. D.; Roux, B.; Watanabe, M., Eds. *Computational Biochemistry and Biophysics*; Marcel-Dekker, Inc.: New York, **2001**.

(3) Neria, E.; Fischer, S.; Karplus, M. *Journal of Chemical Physics* **1996**, *105*, 1902.

(4) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.;

Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *Journal of Physical Chemistry B* **1998**, *102*, 3586.

(5) Jorgensen, W. L.; Tiradorives, J. *Journal of the American Chemical Society* **1988**, *110*, 1657.

(6) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179.

(7) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; BIOMOS b.v.: Zurich, **1996**.

(8) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr Opin Struct Biol* **2009**, *19*, 120.

(9) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophysical Journal* **2008**, *94*, L75.

(10) Shaw, D. E.; Dror, R. O.; K., S. J.; P., G. J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; M., D. M.; B., B.; Bowers, K. J.; Chow, E.; P., E. M.; Ierardi, D. J.; Klepeis, J. L.; S., K. J.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. *Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the 2009 ACM/IEEE Conference on Supercomputing (SC09).* ACM Press: Washington, DC, **2009**.

(11) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Current Opinion in Structural Biology* **2008**, *18*, 134.

(12) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophysical Journal* **2007**, *92*, 3817.

(13) Friesner, R. A. *Adv Protein Chem* **2006**, *72*, 79.

(14) Best, R. B.; Buchete, N. V.; Hummer, G. *Biophysical Journal* **2008**, *95*, L7.

(15) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins-Structure Function and Bioinformatics* **2006**, *65*, 712.

(16) Mackerell, A. D.; Feig, M.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1400.

(17) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins-Structure Function and Bioinformatics* **2010**, *78*, 1950.

(18) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *Journal of Molecular Biology* **1963**, *7*, 95.

(19) Ramachandran, G. N.; Sasisekharan, V. *Adv Protein Chem* **1968**, *23*, 283.

(20) Feig, M. *Journal of Chemical Theory and Computation* **2008**, *4*, 1555.

(21) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *Journal of the American Chemical Society* **1997**, *119*, 5908.

(22) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *Journal of the American Chemical Society* **1984**, *106*, 765.

(23) Headgordon, T.; Headgordon, M.; Frisch, M. J.; Brooks, C. L.; Pople, J. A. *Journal of the American Chemical Society* **1991**, *113*, 5989.

(24) Bohm, H. J. *Journal of the American Chemical Society* **1993**, *115*, 6152.

(25) Gould, I. R.; Cornell, W. D.; Hillier, I. H. *Journal of the American Chemical Society* **1994**, *116*, 9250.

(26) Ono, S.; Kuroda, M.; Higo, J.; Nakajima, N.; Nakamura, H. *Journal of Computational Chemistry* **2002**, *23*, 470.

(27) Cortis, C. M.; Langlois, J. M.; Beachy, M. D.; Friesner, R. A. *Journal of Chemical Physics* **1996**, *105*, 5472.

(28) Grant, J. A.; Williams, R. L.; Scheraga, H. A. *Biopolymers* **1990**, *30*, 929.

(29) Schafer, L.; Newton, S. Q.; Momany, F. A.; Klimkowski, V. J. *Theochem-Journal of Molecular Structure* **1991**, *78*, 275.

(30) Perczel, A.; Angyan, J. G.; Kajtar, M.; Viviani, W.; Rivail, J. L.; Marcoccia, J. F.; Csizmadia, I. G. *Journal of the American Chemical Society* **1991**, *113*, 6256.

(31) Brooks, C. L.; Case, D. A. *Chemical Reviews* **1993**, *93*, 2487.

(32) Dudek, M. J.; Ponder, J. W. *Journal of Computational Chemistry* **1995**, *16*, 791.

(33) Liu, J.; Kelly, C. P.; Goren, A. C.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G.; Zhan, C. G. *J Chem Theory Comput* **2010**, *6*, 1109.

(34) Kang, M.; Smith, P. E. *J Comput Chem* **2006**, *27*, 1477.

(35) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *Journal of the American Chemical Society* **1996**, *118*, 11225.

(36) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins-Structure Function and Genetics* **2000**, *40*, 389.

(37) Maxwell, D. S.; Tiradorives, J.; Jorgensen, W. L. *Journal of Computational Chemistry* **1995**, *16*, 984.

(38) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(39) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J Am Chem Soc* **2007**, *129*, 1179.

(40) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *Journal of Computational Chemistry* **2005**, *26*, 1701.

(41) Lindahl, E.; Hess, B.; van der Spoel, D. *Journal of Molecular Modeling* **2001**, *7*, 306.

(42) Miyamoto, S.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 952.

(43) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *Journal of Chemical Physics* **1995**, *103*, 8577.

(44) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *Journal of Physical Chemistry* **1987**, *91*, 6269.

(45) Hu, H.; Elstner, M.; Hermans, J. *Proteins* **2003**, *50*, 451.

(46) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *Journal of Chemical Theory and Computation* **2007**, *3*, 156.

(47) Koshland, D. E.; Nemethy, G.; Filmer, D. *Biochemistry* **1966**, *5*, 365.

(48) Ding, K.; Gronenborn, A. M. *J Am Chem Soc* **2004**, *126*, 6232.

(49) Hennig, M.; Bermel, W.; Schwalbe, H.; Griesinger, C. *Journal of the American Chemical Society* **2000**, *122*, 6268.

(50) Hu, J. S.; Bax, A. *Journal of the American Chemical Society* **1997**, *119*, 6360.

(51) Wirmer, J.; Schwalbe, H. *Journal of Biomolecular Nmr* **2002**, *23*, 47.

(52) Karplus, M. *Journal of Chemical Physics* **1959**, *30*, 11.

(53) Butterfoss, G. L.; Hermans, J. *Protein Science* **2003**, *12*, 2719.

(54) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 6946.

(55) Best, R. B.; Lindorff-Larsen, K.; DePristo, M. A.; Vendruscolo, M. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 10901.

(56) Marqusee, S.; Baldwin, R. L. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 8898.

(57) Shalongo, W.; Dugad, L.; Stellwagen, E. *Journal of the American Chemical Society* **1994**, *116*, 8288.

(58) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *Journal of the American Chemical Society* **2002**, *124*, 11258.

(59) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *Journal of the American Chemical Society* **2004**, *126*, 7238.

(60) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *Journal of Molecular Biology* **2004**, *336*, 241.

Chapter 5 - Summary and Future Work

Molecular Dynamic (MD) simulations have provided detailed insight into intermolecular and intramolecular interactions in biological systems. Kirkwood Buff theory provides a direct connection between the particle distributions and thermodynamic properties. Additionally, KB theory can be applied to both experimental data and simulation result. The combination of KB theory and MD simulation can be performed to multiple-component systems to reproduce a series of physical properties and disclosure the atomic behavior. As the improvement of backbone torsional potential, the precision of peptide and protein simulation boost dramatically, indicating a higher level extension on the range of applicable systems is available from MD simulation.

The Smith group has been working on developing Kirkwood Buff force field to biological systems for several years and now we are almost close. We believe in the near future the whole set of KBFF will be completed and lead us to more accurate simulation towards discovering protein folding mechanism.

Appendix A - A Kirkwood-Buff Derived Force Field for Aqueous Alkali Halides

Moon Bae Gee, Nicholas R. Cox, Yuanfang Jiao, Nikolaos Bentenitis, Samantha Weerasinghe, and Paul E. Smith. *Journal of Chemical Theory and Computation*, **2011**, 7, 1369-1380

A.1 Abstract

A classical nonpolarizable force field is presented for the simulation of aqueous alkali halide solutions (MX), where M=Li^+, Na^+, K^+, Rb^+, and Cs^+ and X = F^-, Cl^-, Br^-, and I^-, and their interactions with biomolecules. The models are specifically designed to reproduce the experimental Kirkwood-Buff integrals, and thereby the solution salt activities, as a function of salt concentration. Additionally, we demonstrate that these models reasonably reproduce other experimental properties including ion diffusion constants, dielectric decrements, and the excess heats of mixing. The parameters are developed by considering the properties of aqueous NaX and MCl solutions using a previously established model for NaCl. Transferability of the parameters to other salts is then established by the successful simulation of additional aqueous salt solutions, KI and CsBr, not originally included in the parametrization procedure.

A.2 Introduction

Aqueous solutions of alkali metal halides are not only the simplest models for the aqueous electrolyte solutions but also play an important role in many biological systems. They can help to stabilize biomolecules, such as proteins, nucleic acids, and lipids, and are often involved in biological catalysis.[1-3] Because of their importance in biological phenomena, and the desire to study these more complicated ternary systems using computer simulation, many force fields for alkali metal cations and halide anions have been reported in the literature.[4-11] A recent comprehensive survey has also been provided by Joung and Cheatham.[4] The wide range of parameter sets available for salt systems is, in our opinion, a direct result of the fact that there is relatively little experimental data available that is both sensitive to changes in the ion parameters and also easily amenable to simulation. Furthermore, as our ability to access longer simulation time scales has improved, a number of problems with many of the existing ion force fields have recently come to light.[12,13] One approach to solving these problems is the use of models which

explicitly include polarization effects.[14-16] However, as this significantly increases the computational demand, the vast majority of biomolecular simulations still do not include explicit polarization effects. Therefore, there remains a need for simple but reliable ion force fields, especially for systems displaying slow relaxation times.

Recently, there have been three major attempts to develop force fields for all alkali metals and halide ions. Jensen and Jorgensen have developed TIP4P water compatible alkali halide parameters using the ion hydration free energies and ion-water contact distances as target data.[11] Joung and Cheatham[4] have also used the free energy of hydration for individual ions, as well as the lattice energies and the lattice constants of alkali metal halides and gas phase ion-water interaction energies, in order to produce force fields for all of the alkali metal and halide ions which are compatible with three commonly used nonpolarizable water models, namely, SPC/E, TIP3P, and TIP4P$_{EW}$. Horinek et al.[17] have used both the free energy and the entropy of hydration of the individual ions in order to parametrize their force fields and focused on the nonpolarizable SPC/E water model. Horinek et al. argued that their force field would be more applicable for biomolecular simulations where the salt concentrations are low, while the Joung and Cheatham force fields would be more applicable when the salt concentrations are moderate. All three force fields attempt to reproduce a series of initial properties, including the free energies (and entropies) of hydration, the first peak of the ion-water radial distribution function (rdf), gas phase ion-water binding energies, and crystal lattice parameters. However, they were essentially developed using properties that that do not directly probe ion-ion interactions in solution. A subsequent study has since evaluated the solute activity for two salts using the Joung and Cheatham force fields obtained using thermodynamic integration.[18] This does probe ion-ion interactions. However, the study provided only moderate success—good results were obtained for KCl, but significant deviations from experimental results were observed for NaCl solutions above 0.5 m.[18] The comparison of simulated and experimental diffusion constants and solubilities also provided mixed results.

We have taken a very different approach in an attempt to develop accurate force fields for solution mixtures. Our approach is based on the thermodynamics of solution mixtures as described by Kirkwood-Buff (KB) theory.[19-26] Here, the central properties of interest are the Kirkwood-Buff integrals (KBIs) defined by

$$G_{ij} = 4\pi\int_0^\infty [g_{ij}^{\mu VT}(r) - 1]r^2 dr$$

(A.1)

where G_{ij} is the KBI between species *i* and *j*, $g_{ij}^{\mu VT}(r)$ is the corresponding radial distribution function (rdf) in the grand canonical ensemble at the composition of interest, and r is the center of mass distance between the two species. An excess coordination number can be defined by $N_{ij} = \rho_j G_{ij}$, where $\rho_j = N_j/V$ is the number density of *j* particles. The physical meaning of the excess coordination number is the difference in the number of *j* species in the vicinity of a central *i* species on the addition of the *i* species from that found in an equivalent volume of bulk solution. Hence, a value of N_{ij} significantly greater than zero indicates an excess of species *j* in the vicinity of species *i* (over the random bulk distribution), while a significant negative value corresponds to a depletion of species *j* surrounding *i*. Combinations of KBIs provide expressions for a variety of thermodynamic properties of the solution of interest.[27,28]

Kirkwood-Buff theory can then be used to relate solution structure, in terms of the KBIs, to the thermodynamic behavior of the solution.[29-31] The expressions provided by KB theory are exact, and the theory involves no approximations beyond the usual statistical mechanical assumptions (larger number of molecules, thermodynamic limit, etc). The expressions can be applied to study any stable solution mixture involving any number of components of any type (small molecules through to proteins) at any composition and any temperature and pressure. The analysis of experimental data for solution mixtures using KB theory is well established and provides quantitative information concerning species distributions in solutions and how they vary with composition.[28,29,32] The resulting KBIs can also be obtained from computer simulations and thereby provide a rigorous test of the accuracy of current force fields.[31,33]

Our parameters were developed to reproduce the properties of solution mixtures and are therefore collectively known as Kirkwood-Buff derived force fields (KBFF).[19-25] The parameters for the KBFF models are determined using a combination of molecular dynamics simulation, the Kirkwood-Buff (KB) theory of solutions, and available experimental data concerning activity coefficients and solution densities. This approach has several advantages. First, KB theory is exact and includes no approximations. Second, KB theory can be applied to any stable solution mixture. Third, the KB integrals are easily obtained from the radial distribution functions (rdf) provided by MD simulations and are quite sensitive to the force field parameters. Fourth, the KB integrals help quantify the distributions arising from the relative strengths of the solute_solute

and solute_solvent interactions.[25,34] Hence, the general philosophy of the Kirkwood-Buff derived force field approach is to use the KBIs obtained from an analysis of the experimental data as target values for the development of accurate force fields for a variety of solutes. The target data are composition-dependent, and this dependence is also used during the parametrization process. We have argued that reasonable agreement with experimental results is also obtained for other properties not included in the original parametrization.[19,20,22-25] In doing so, we favor the use of data for solution mixtures, primarily the KBIs, and are less influenced by gas phase data or infinite dilution data such as free energies of hydration. A model for aqueous NaCl solutions has already been developed using this type of approach,[25] and here we simply generalize this initial model to include other alkali halide salts.

Recently, two research groups also produced KB derived force fields for some of the alkali metal halides. Hess and van der Vegt used the SPC/E water model to develop KB-derived force fields for Li^+ and K^+ in order to explain the differential binding affinity of alkali metal ions to carboxylate ions.[35] Klasczyk and Knecht used the SPC water model and the KBFF force field for the chloride ion to develop force fields for Li^+, K^+, Rb^+, and Cs^+, but not for halide ions.[36] Therefore, the more extensive Klasczyk and Knecht force field is incomplete and, in principle, incompatible with our models because we use the SPC/E water model. In this paper, we present aKB derived force field for a wide variety of alkali metal and halide ions. The models are intended to be applicable over the whole concentration range and are consistent with our previous models for a variety of solutes in both aqueous and nonaqueous solutions.

A.3 Methods

A.3.1 Kirkwood-Buff Analysis of Alkali Halide Solutions

The complete details concerning the extraction of the KBIs from the experimental data, the so-called Kirkwood-Buff inversion procedure, have been provided elsewhere.[27,28,37] For a binary solution consisting of water (w) and a salt cosolvent (c), a variety of thermodynamic quantities can be defined in terms of the KB integrals G_{ww}, G_{cc}, and $G_{cw} = G_{wc}$ and the number densities (or molar concentrations) ρ_w and ρ_c.[25] By use of the KB inversion procedure, one can also extract the composition-dependent KBIs from the corresponding experimental thermodynamic properties.[28] Specifically, the KB inversion approach uses composition-dependent experimental binary solution data for the isothermal compressibility, partial molar

volumes, and cosolvent activity in order to extract the corresponding three KBIs using the expressions provided by KB theory. Hence, KB theory provides a link between measurable experimental data and the species distributions in solution, which are then quantified in terms of the KBIs. The relationships used for the present work are[27]

$$1 + N_{cc} = \rho_c RT\kappa_T + \rho_w^2 \frac{\overline{V_w}^2}{\mu_{cc}}$$
$$1 + N_{ww} = \rho_w RT\kappa_T + \rho_w \rho_c \frac{\overline{V_c}^2}{\mu_{cc}}$$
$$N_{wc} = \rho_c RT\kappa_T - \rho_w \rho_c \frac{\overline{V_w}\,\overline{V_c}}{\mu_{cc}}$$

(A.2)

where κ_T is the isothermal compressibility, V_i are partial molar volumes, and μ_{cc} represents a chemical potential (or activity) derivative given by

$$\mu_{cc} = \beta\left(\frac{\partial \mu_c}{\partial \ln m_c}\right)_{T,P} = 1 + \left(\frac{\partial \ln \gamma_c}{\partial \ln m_c}\right)_{T,P}$$

(A.3)

where $\gamma_c = \gamma_\pm$ is the molar activity coefficient of the salt and mi is the molality of *i*. Hence, the three KBIs can be obtained from a knowledge of the compressibility, partial molar volumes (or density), and activity as a function of the composition (three equations in three unknowns).

Experimental activity coefficient data at 298.15 K and 1 atm were taken from the literature,[38] and fitted to the following functional form,[38,39]

$$\ln \gamma_\pm = -\frac{1.18\sqrt{m_s}}{1 + a_1\sqrt{m_s}} - \ln\left(1 - a_2 m_s\right) + a_3 m_s + a_4 m_s^2$$

(A.4)

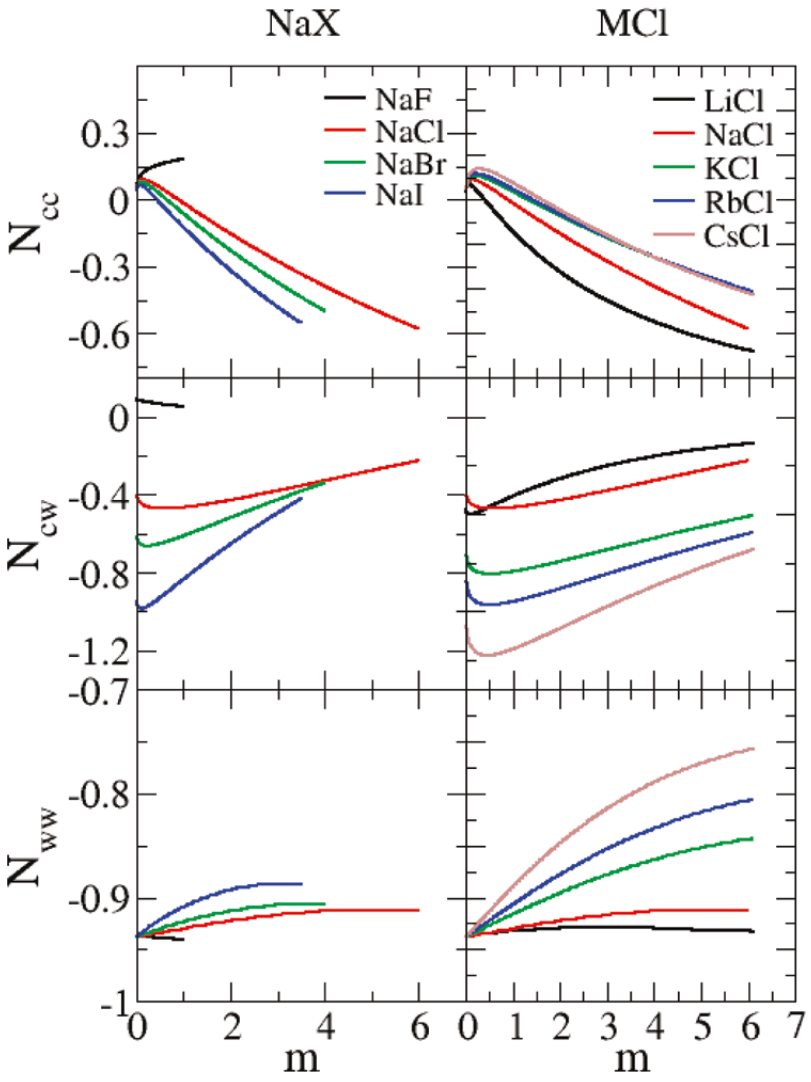
where m_s is the salt molality and the *a*'s represent fitting parameters with no particular physical meaning. The first term on the right-hand side of eq A.4 is a Debye-Hueckel term for 1:1 salts which is required to fully capture the correct behavior of salts at low salt concentrations. Issues

associated with the quality of fit for the experimental activity coefficient data provide the main source of error in the KB analysis. The final fitting parameters are provided in the Supporting Information. Previously established polynomial fitting expressions for the experimental density data of salts[40] were used to determine partial molar volumes using standard approaches.[41] The solution compressibility has a negligible effect on the resulting KBI values for solutions at moderate temperatures and pressure.[32] Hence, the compressibility was assumed to follow the simple relationship, $\kappa_T = \varphi_w \kappa^0_{Tw} + \varphi_c \kappa^0_{Tc}$, where φ_i is the volume fraction and κ^0_{Ti} is the compressibility of the pure substance (water or salt). The compressibility of pure water was taken to be 4.6 x 10^{-10} m^2/N,[42] while the compressibilities of the salt crystals were taken to be zero. The experimental compressibility (approximated), partial molar volumes, and activity provided by eq 4 were then used with the expressions provided in eq A.2 to isolate the experimental KBIs as a function of the composition. The results of the KB inversion analysis are presented in Figure A.1.

A.3.2 Kirkwood-Buff Theory of Salt Solutions

Some complications arise when applying KB theory to salt solutions.[25,43] First, the salt can dissociate into free cations and anions (we will assume complete dissociation for the salts examined here). Second, electroneutrality constraints for regions of the solution surrounding each species provide additional relationships between the KBIs.[43] Let us consider a salt containing a total of n ions which will fully dissociate to provide nt cations and n_ anions. If one chooses the salt as the relevant thermodynamic species, then $d\mu_s = nRTdln(m_s\gamma_\pm)$ and the activity derivatives provide a set of KBIs (G_{ss} and G_{sw}) involving the salt "molecules" when using the KB inversion approach. However, this choice is rather awkward from the simulation point of view as we typically observe free ions for strong electrolytes, and therefore the rdf's between salt "molecules" are difficult, if not impossible, to determine. Consequently, in this work, the salt solution is treated as a binary system of indistinguishable ions (c) and water (w), and we will distinguish between the cosolvent (total ion) concentration, m_c or ρ_c, and the classic salt concentration, m_s or ρ_s. Consequently, for a n_+:n_- salt, one has $nm_s = m_c$, $n\rho_s = \rho_c$, $V_s = nV_c$, and $\gamma_c = \gamma_\pm$. In addition, the following relationships are also obeyed: $d\mu_s = nd\mu_c$, $\rho_s d\mu_s = \rho_c d\mu_c$, $d\ ln\ m_s = d\ ln\ m_c$, $\rho_s V_s + \rho_w V_w = \rho_c V_c + \rho_w V_w = 1$, and $\rho_c d\ ln\ a_c = \rho_w d\ ln\ a_w = \rho_s d\ ln\ a_s + \rho_w d\ ln\ a_w = 0$, at constant p and T—the latter being the Gibbs-Duhem equation.

Figure A.1 Experimentally derived excess coordination numbers for aqueous alkali halide solutions as a function of salt molality at 298.15 K and 1 atm.



Hence, the experimental data can then be analyzed in terms of either salt molecules or a collection of indistinguishable ions. The resulting KBIs obtained from the two formalisms are related by

$$G_{ss} = \frac{1-n}{\rho_c} + G_{cc} \qquad G_{sw} = G_{cw} \tag{A.5}$$

The KBIs obtained from the indistinguishable ion approach (G_{cc} and G_{cw}) involve rdf's between the ions (and water molecules), which ignore the ion identity (cation or anion). The relationships between the KBIs using the cosolvent label and those involving the anion/cation label are provided by

$$G_{cc} = \left(\frac{n_+}{n}\right)^2 G_{++} + \left(\frac{n_-}{n}\right)^2 G_{--} + \frac{n_+ n_-}{n^2}\left(G_{+-} + G_{-+}\right)$$
$$G_{cw} = G_{wc} = \frac{n_+}{n} G_{+w} + \frac{n_-}{n} G_{-w}$$

(A.6)

and were obtained in a similar manner as done previously.[25] Here, the KBI denoted as G_{++} refers to the integral over the cation-cation rdf in solution. We note that the above relationships merely reflect a change in indices and do not invoke the electroneutrality conditions.

If one then assumes that electroneutrality must be obeyed in the local regions surrounding each molecule or ion,[22,25,43] then one can show that the following relationships must also hold:

$$G_{cc} = -\frac{1}{\rho_c} + G_{+-} \qquad G_{cw} = G_{+w} = G_{-w}$$
$$G_{+-} = \frac{1}{\rho_+} + G_{++}$$
$$\frac{1}{\rho_+} + G_{++} = \frac{1}{\rho_-} + G_{--}$$

(A.7)

where ρ_+ is the number density of cations etc. Hence, all of the ion-ion KBIs are related, and there is only one independent KBI for a binary solution. We choose this to be G_{cc} for the present analysis.

A.3.3 Molecular Dynamics Simulations

All molecular dynamics simulations of alkali halide solutions were performed using the SPC/E water model[44] in the isothermal isobaric (*NpT*) ensemble at 300 K and 1 atm as implemented in the GROMACS program (v3.3.1).[45,46] A time step of 2 fs was used, and the geometry of the water molecules was constrained using SETTLE.[47] The weak coupling technique was used to modulate the temperature and pressure with relaxation times of 0.1 and

0.5 ps, respectively.[48] The particle mesh Ewald technique (PME) was used to evaluate electrostatic interactions using a cubic interpolation and a grid spacing of 0.1 nm for the reciprocal space sum, coupled with tinfoil boundary conditions.[49] The initial cubic boxes for each solution at the required concentration were generated by randomly placing water molecules with ions starting from pure solvent boxes of length varying between 4 and 6 nm. During the simulations, configurations were saved every 0.1 ps for analysis. Diffusion constants were determined using the mean square fluctuation approach,[50,51] and relative permittivities were obtained from the dipole moment fluctuations.[52,53] The excess enthalpy of mixing (ΔH_{mix}) was determined using an established procedure which uses the average potential energies[54] and the configurational energies from the pure SPC/E water and the alkali halide lattice.

A.3.4 Kirkwood-Buff Analysis of the Simulation Data

Radial distribution functions were obtained for each system and composition. The pair rdf's thereby correspond to the ion-ion, ion-water, and water-water distributions after averaging over all other ions and water molecules at that particular composition. The indistinguishable ion treatment for salts involves the determination of ion-ion and ion-water rdf's, which ignore the identity of the ions involved. For example, in NaCl solutions, the ion-water rdf is determined after averaging over the ion-water distributions using both types of ion, sodium and chloride, at the origin. The Kirkwood-Buff integrals (KBIs) are defined for systems open to all the solution components. However, most simulations are performed in closed systems. Hence, one has to approximate the KBIs by truncating the integral after a certain distance

$$G_{ij} \approx 4\pi \int_0^R [g_{ij}^{NpT}(r) - 1] r^2 dr \qquad \text{(A.8)}$$

where R represents a correlation distance within which the solution composition differs from the bulk composition. This approximation has been shown to be very reasonable as long as the systems are not too small (L > 4 nm) and sufficient sampling (> 5 ns) is achieved.[26,29,55] The values of G_{ij} used here were determined by averaging the integral over a short-range of distances (1.2 - 1.5 nm), taken as approximately one water-water solvation shell. The final values were relatively insensitive to the exact distance and range used, but this approach helps to reduce

statistical fluctuations associated with the integrals. Once the three simulated KBIs have been obtained from the trajectory at a particular bulk composition, one can then use these values in a series of expressions which provide thermodynamic properties of the solution mixture. The partial molar volumes of the components (V_i) are given by[41]

$$\overline{V_w} = \frac{1+\rho_c(G_{cc}-G_{cw})}{\eta} \qquad \overline{V_c} = \frac{1+\rho_c(G_{ww}-G_{cw})}{\eta},$$
$$\eta = \rho_w + \rho_c + \rho_w\rho_c(G_{ww}+G_{cc}-2G_{cw})$$

(A.9)

Using the simulated KBIs, one can determine a variety of derivatives of the chemical potential, depending on the concentration scale used. Here, we choose derivatives of the activity with respect to molarity.[25] Of primary interest is the following activity derivative:

$$a_{cc} = \left(\frac{\partial \ln a_c}{\partial \ln \rho_c}\right)_{p,T} = 1 + \left(\frac{\partial \ln y_c}{\partial \ln \rho_c}\right)_{p,T} = \frac{1}{1+\rho_c(G_{cc}-G_{cw})}$$

(A.10)

where a_c and y_c are the cosolvent (average ion) molar activity and molar activity coefficient, respectively. Hence, changes in the cosolvent activity can be determined directly from the simulations. Furthermore, accurate activity derivatives ensure reasonable activities are thereby obtained. The partial molar volumes and activities obtained in this manner have been shown to be in agreement with the results obtained using alternative computational approaches.[21,56]

A.3.5 Parameter Development

The KBFF models used in this study involve a simple classical nonpolarizable description for each molecule. The intermolecular interactions are described by the Coulomb and Lennard-Jones (LJ) 6-12 potentials, which contain just two adjustable parameters for ions, namely, the Lennard-Jones diameter (σ) and the interaction strength (ε). In this scheme, each pair of atoms *i* and *j* interact with an interaction energy given by

$$V_{ij} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

(A.11)

Here, all of the symbols have their usual meaning.[1] This model was chosen so as to be computationally efficient, while maintaining compatibility with existing force fields and programs used for the simulation of biomolecules. The ion parameters are combined with the SPC/E model for water.[44] Geometric combination rules were used for both σ and ε. In order to obtain parameters for the LJ term, we have employed the same method described previously for NaCl.[25] This approach requires three pieces of experimental data: ionic radii of alkali and halide ions that are consistent with the crystal lattice dimensions, crystal lattice unit cell dimensions, and the ion to water oxygen contact distances (see Table 1). These data were chosen in an effort to be both consistent with our previous force fields and to help restrict the range of possible σ and ε values to be studied. However, satisfactory agreement with the experimental data was not possible for all ions using this simple approach (see below).

Table A.1 Experimental Data Used during the Initial Parameter Development[a]

	MCl					NaX			
	Li^+	Na^+	K^+	Rb^+	Cs^+	F^-	Cl^-	Br^-	I^-
r (nm)	0.115	0.101	0.138	0.149	0.170	0.133	0.181	0.196	0.220
a (nm)	0.257	0.282	0.319	0.332	0.412	0.239	0.282	0.299	0.324
d (nm)	0.213	0.240	0.280	0.289	0.314	0.263	0.319	0.338	0.365
ref	60, 62	25	60, 62	60, 62	60, 62	60, 62	25	60, 62	60, 62

[a] *r*, the ionic radii of alkali halide ions which are consistent with the crystal lattice dimensions; *a*, the crystal lattice unit cell dimension; and *d*, the ion to water oxygen contact distance.

The first step was to parametrize the anions (F^-, Br^-, I^-) by studying the crystal structures and several aqueous solutions of NaF, NaBr, and NaI, using the same Nat parameters from our previous NaCl study.[25] The values of σ_{--} were determined by scaling the ionic radii of each ion with the same scaling factor as used previously (2.43).[25] The values of ε_{--} were then varied in an effort to reproduce the experimental lattice dimensions of the sodium halide crystals, and the anion-water contact distances, in the simulations. The final values determined for each ion were then used to provide the simulated KBIs for a variety of aqueous solutions. Unfortunately, in the case of the F^- anion, a reasonable value for σ_{FF} which reproduced the crystal lattice dimensions

could not be obtained by a simple scaling approach. Hence, we decided to develop specific values of σ_{FF} (and ε_{FF}), which attempted to reproduce both the crystal lattice dimensions and solution KBIs.

Table A.2 Final Force Field Parameters Describing the KBFF Models for Alkali Halides[a]

model	atom	σ_{ii} (nm)	ε_{ii} (kJ/mol)	σ_{iO} (kJ/mol)	q (e)
KBFF	Li	0.1820	0.7000	0.2700	+1.0
	Na	0.2450	0.3200	0.3420	+1.0
	K	0.3340	0.1300	0.2327	+1.0
	Rb	0.3620	0.1500	0.2655	+1.0
	Cs	0.4130	0.0065	0.1954	+1.0
	F	0.3700	1.0000		-1.0
	Cl	0.4400	0.4700		-1.0
	Br	0.4760	0.3000		-1.0
	I	0.5350	0.2000		-1.0
SPC/E	O	0.3166	0.6506		-0.8476
	H	0.0000	0.0000		+0.4238

[a] The following combination rules used: $\sigma_{ij} = (\sigma_{ii} \times \sigma_{ij})^{1/2}$, $\varepsilon_{ij} = s(\varepsilon_{ii} \times \varepsilon_{ij})^{1/2}$. The value of s was set to unity for all interactions except for cation to water oxygen, where values of s = 0.4 (Li), 0.75 (Na), 0.8 (K), 0.85 (Rb), and 0.95 (Cs) were used. The NaCl ion and SPC/E water parameters were taken from previous studies.[25,44]

Second, the initial cation parameters for Li^+, K^+, Rb^+, and Cs^+ were developed by reference to the crystal dimensions of LiCl, KCl, RbCl, and CsCl and the relevant cation-water contact distances. After the values of σ_{++} were determined by scaling the ionic radii of each ion, the values of ε_{++} were varied to reproduce the crystal unit cell dimensions and the cation-water contact distances. Unfortunately, and in agreement with our earlier study of NaCl,[25] we could not reproduce the experimental KBIs in aqueous solution by using standard combination rules for ε_{++} in aqueous solutions. Hence, modified ε parameters were developed specifically for the cation-water oxygen interactions. This interaction was subsequently modified by introducing a simple scale factor (s) for the interaction between metal ions and water oxygens such that $\varepsilon_{MO} = s(\varepsilon_{MM}$

$\varepsilon_{OO})^{0.5}$. This parameter scales the repulsive part of the LJ potential controlling the contact distance between an ion and first shell water molecules. The scale factor was set to unity for all other interactions. The final scaling factors for the metal ion-water interactions are provided in Table 2. Unfortunately, this simple approach did not work for LiCl. Hence, unique (not scaled) LJ values were determined for this salt by reference to the LiCl crystal dimensions and solution KBIs.

A.4 Results

The main goal for the force fields developed here is to reproduce, as far as possible, the experimental KBIs for aqueous salt solutions as a function of salt concentration. Hence, we present this comparison first. This is followed by a comparison of a series of additional properties of solution mixtures, not included in the original parametrization, which is presented in an effort to both fully characterize the models and to establish the range of applicability of the models. As the solutions involve a variety of highly polarizing ions, the inherent many body interactions would be expected to vary substantially between different salts and also with concentration. Therefore, it should be obvious that it is essentially impossible to reproduce all the available experimental data using such a simple LJ 6-12 plus Coulomb model. Wherever possible, we have attempted to highlight any disagreement with experimental results and possible causes for these errors.

The experimental excess coordination numbers for sodium halides and alkali chlorides are displayed in Figure A.1. The results presented in Figure 1 have been extracted from the experimental thermodynamic data on aqueous salt solutions and represent the primary target data for the current parametrization approach. The data display systematic trends between the different salts, which provide information concerning the underlying molecular distributions. At low concentrations (<0.1 m), the distributions are dominated by the Debye-Hueckel behavior leading to positive values for the ion-ion excess coordination numbers (N_{cc}). This behavior reverses at higher salt concentrations and indicates, with the exception of NaF, an increase in ion solvation by water. Similar results have been observed in other studies.[57,58]

Table A.2 shows the final Lennard-Jones parameters used in our simulations. The LJ parameters for Na^+ and Cl^- were taken from Weerasinghe and Smith.[25] As the size of the cation increased, the value of σ increased and that of ε essentially decreased. A similar trend is observed

for the anions. Peng et al. have argued in favor of such trends in the LJ parameters, although the trend in ε parameters is the opposite of that expected (decreasing with atomic number, not increasing).[6] Their work used a LJ 9-12 potential, and hence the argument might not be so clear for the LJ 6-12 plus Coulomb models, or for systems with large polarization effects, where the ε parameter is linked to a scaling of the repulsive wall which resists the electrostatic attraction, rather than the usual relationship to dispersion interactions. The trend in the values of σ was also observed by both Joung and Cheatham4 and Horinek et al.[17] However, any trend in the values of ε was absent from both these previous works.

Table A.3 indicates the potential energy, density, and lattice constants obtained for the salt crystals studied in this work. The simulated crystal dimensions exhibit an average error of 3% with a maximum error of 10%. In the Supporting Information, the lattice energies of the Kirkwood-Buff models are compared to the experimental data and the force fields developed by Peng et al.[6] The KBFF models consistently overestimate the lattice energies. While reproducing the crystal lattice energies of salts was not a goal of the present parametrization, the results suggest that the current force fields may result in crystal lattices which are too stable with respect to the solution phase. This could be a concern for future simulations. However, a recent study of the KBFF model for NaCl indicates an approximate solubility of 7.9 m,[59] compared to the experimental value of 6.1 m.[60] The higher observed solubility suggests that, if anything, the opposite could be true. Some of these differences are probably related to the rather crude LJ 6-12 potential used in the current work which is known to fail for crystals.[6] Our main aim in studying the salt crystal lattice properties was to guide the systematic development of anion and cation LJ σ parameters. Furthermore, the enthalpies of mixing appear to be well reproduced (see later), indicating good compatibility with the SPC/E water model. Hence, we have not considered any further attempts to significantly improve the current data.

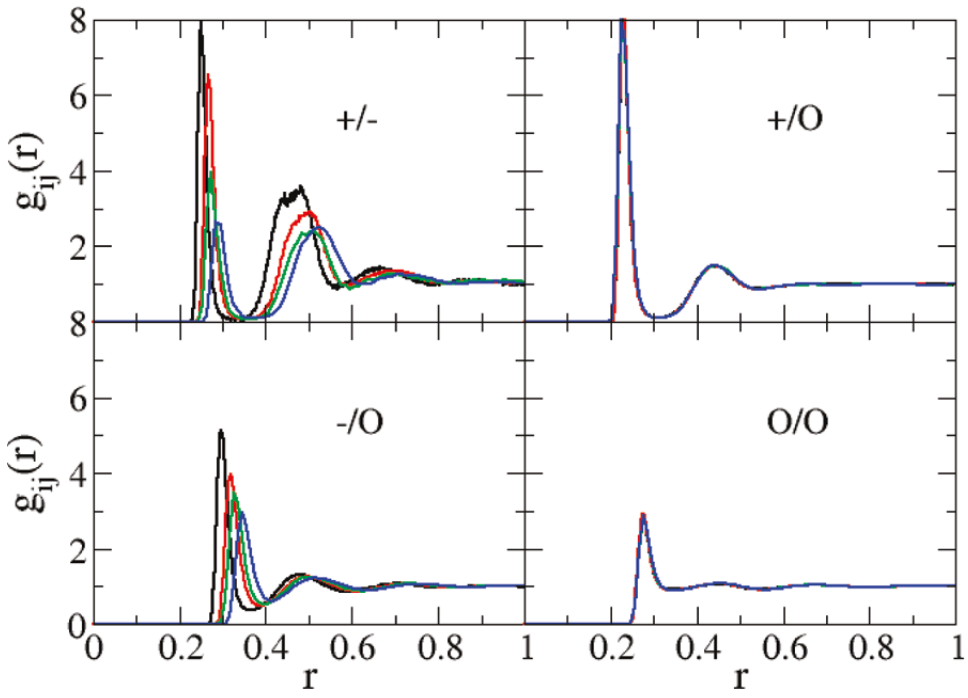
Table A.3 Summary of the Alkali Halide Crystal Simulations Using the Final Parameters[a]

	E_{pot} (kJ/mol)	ρ_{sim} (g/cm^3)	ρ_{exp} (g/cm^3)	a_{sim} (nm)	a_{exp} (nm)
NaF	-1217.74	1.965	2.558	0.257	0.231
NaCl	-808.24	2.108	2.163	0.285	0.281
NaBr	-776.08	3.326	3.246	0.295	0.297
NaI	-750.94	3.878	3.665	0.303	0.323

LiCl	-1178.03	1.776	2.069	0.270	0.257
Kcl	-725.29	1.980	1.990	0.315	0.314
RbCl	-692.73	2.800	2.859	0.325	0.327
CsCl	-650.12	3.990	3.973	0.419	0.412
KI	-663.23	3.406	3.125	0.343	0.353
CsBr	-628.80	4.582	4.453	0.433	0.429

[a] Symbols are Epot, average total potential energy per molecule (Ns); F, mass density; and a, unit cell dimension. Subscripts sim and exp indicate simulation and experimental data,[70] respectively.

Figure A.2 Radial distribution functions obtained from simulations of 1 M sodium salt solutions containing NaF (black lines), NaCl (red lines), NaBr (green lines), and NaI (blue lines). Cations, anions, and the water oxygen are denoted by the symbols +, -, and 0, respectively.



The radial distribution functions (rdf's) obtained from the 1M salt simulations are displayed in Figure A.2 for the sodium halides and in Figure A.3 for the alkali metal chlorides.

The sodium to halide anion-cation rdf's displayed a large first (ion pair) and a significant second (solvent separated ion pair) peak, which is in agreement with experimental results.[61] All rdf's approached unity beyond 1 nm. The first shell coordination numbers, n_{ij}, as well as the distances to the first rdf maximum (contact distance), Rmax, and the first rdf minimum (first solvation shell), Rmin, were calculated from the corresponding rdf's as a function of the solution molality and are presented in the Supporting Information. The final contact distances for Li^+, Na^+, K^+, Rb^+, Cs^+, F^-, Cl^-, Br^-, and I^- were 0.19, 0.23, 0.26, 0.28, 0.29, 0.27, 0.32, 0.33, and 0.35 nm, respectively. As expected, the radius of the first hydration shell increased as the size of the cation and anion increased. The simulated contact distances agree with the experimental values of 0.20, 0.24, 0.28, 0.29, 0.31, 0.26, 0.32, 0.34, and 0.36,62 respectively, to within a 0.01 nm root-meansquare (rms) deviation—a similar deviation to that exhibited by the force field of Joung and Cheatham.[4] The first water shell coordination numbers of Na^+, K^+, Rb^+, and Cs^+ in ~4 M aqueous solutions were determined to be 4.9, 5.9, 6.2, and 6.4, respectively. As expected, and similar to the trend in the radii of the first hydration shell, the hydration numbers increase as the size of the cation increases. The predicted hydration numbers agree with those determined from X-ray and neutron scattering data under the same conditions[61]—4.9, 5.3, 6.9, and 7.5, respectively—to within a 0.2 rms deviation. The Supporting Information also indicates that the coordination numbers are sensitive not only to the size of the alkali metal ion but also to changes in the salt concentration. The degree of ion pairing increases with increasing concentration. We note that no aggregation or crystallization was observed during any of the simulations.

The simulated and experimental excess coordination numbers, N_{ij}, are shown in Figure A.4 for the sodium halides and in Figure A.5 for the alkali metal chlorides, as a function of salt molality. The KBFF models quantitatively reproduce the experimental data, although the simulated values were somewhat less accurate for NaI and CsCl solutions. The correct trends (with salt concentration) are reproduced for all salts. The ion-ion excess coordination numbers (black lines) did not vary significantly from salt to salt when compared to the variation in the ion-water excess coordination numbers (red lines), which is in agreement with the experimental data (see Figure A.1). This suggests that changes to the ion-water and water-water distributions determine the solution behavior to a large extent. However, it is very difficult to clearly relate these composition-dependent changes to the force field parameters used here. The relatively poor agreement for the NaI and CsCl solutions probably arises due to the high polarizability of the

anion and cations, respectively, which would make the development of parameters suitable for both crystals and aqueous solutions quite challenging.

Figure A.3 Radial distribution functions obtained from simulations of 1 M chloride salt solutions containing LiCl (black lines), NaCl (red lines), KCl (green lines), RbCl (blue lines), and CsCl (brown lines). Cations, anions, and the water oxygen are denoted by the symbols +, -, and 0, respectively.

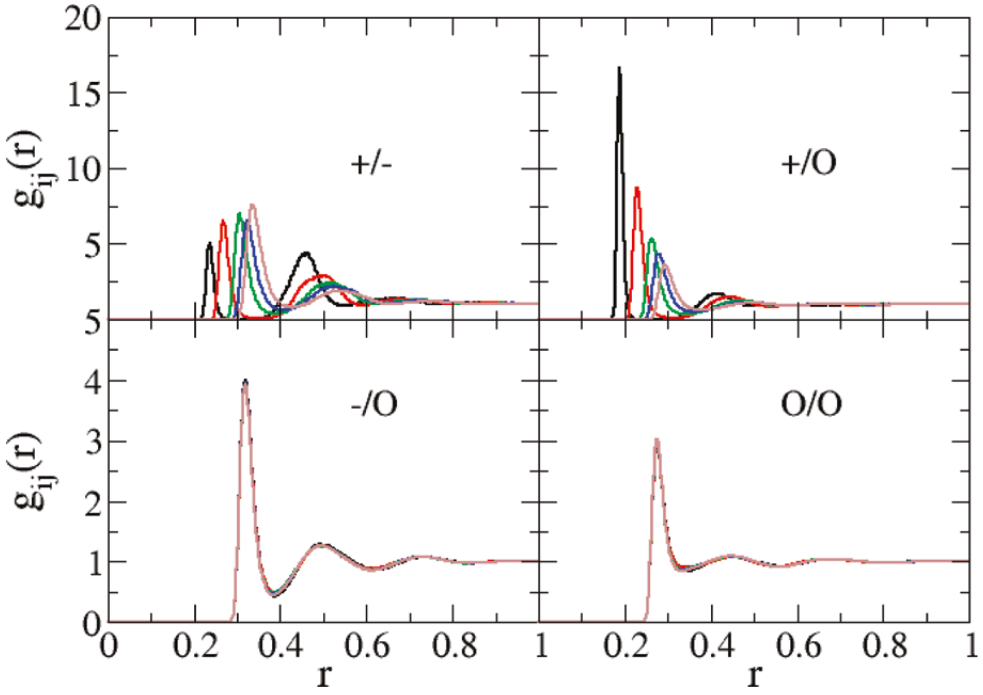
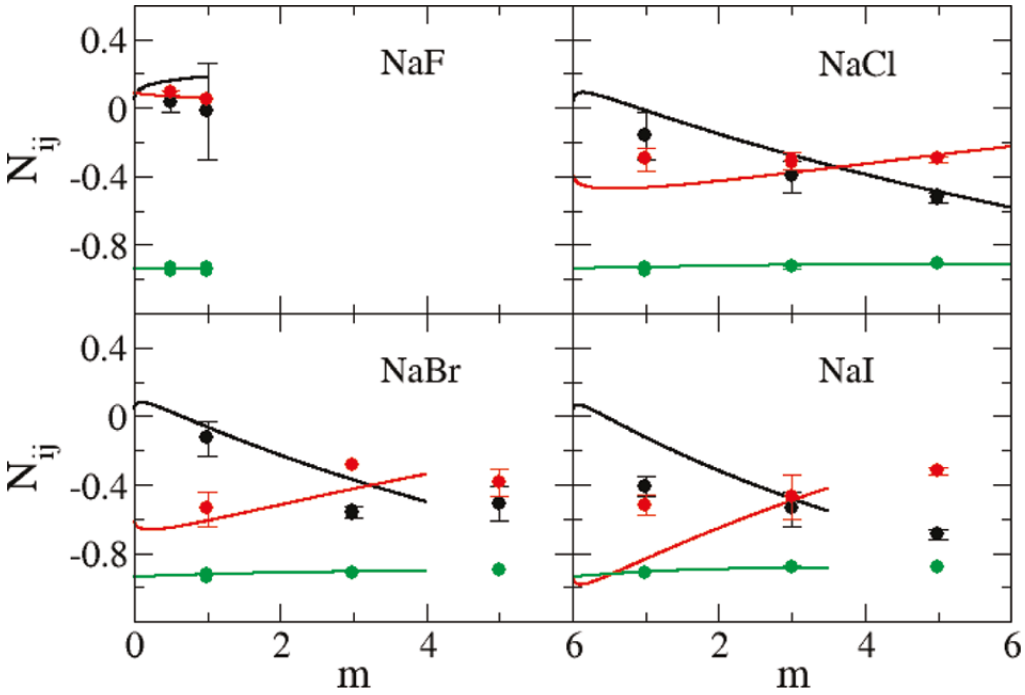


Figure A.4 Excess coordination numbers as a function of salt molality. The N_{cc} (black lines), N_{cw} (red lines), and N_{ww} (green lines) are obtained from a KB analysis of the experimental data. The N_{cc} (black dots), N_{cw} (red dots), and N_{ww} (green dots) are obtained from simulations performed with the KBFF models.



In Figures A.6 and A.7, the simulated activity derivatives (a_{cc}) as a function of molality are compared to the experimental values.[38] The KBFF model reproduced the correct increase in acc with concentrations at higher salt concentrations as indicated by the experimental data. We note that acc plays an important role for solutions as it characterizes the change in activity (chemical potential) of the salt with concentration.[31] Hence, accurate force fields are required to reproduce this data.[25] An expression for the molar activity coefficient ($y_c = y_\pm$) provided by the current force fields was obtained by taking appropriate derivatives of the fitting equations adopted for the experimental data (eq 4) and then obtaining parameters that best fit the simulated activity derivatives. The final fitting parameters are provided in the Supporting Information for most of the salt solutions studied here. It should be noted that many common force fields do not correctly reproduce the experimental excess coordination numbers and activity

derivatives.[20,22,23,25] For instance, in our previous work, we simulated 2 M NaCl solutions using a variety of salt force fields.[25] Many force fields provided values of $a_{cc} < 0.5$. Large deviations from experimental results are also observed for other solutes.[20,34,63] Hence, the data provided in Figures A.6 and A.7 for the present models, while not perfect, can be considered to be in good agreement with experimental results relative to typical results for similar force fields.

Figure A.5 Excess coordination numbers as a function of salt molality. The N_{cc} (black lines), N_{cw} (red lines), and N_{ww} (green lines) are obtained from a KB analysis of the experimental data. The N_{cc} (black dots), N_{cw} (red dots), and N_{ww} (green dots) are obtained from simulations performed with the KBFF models.

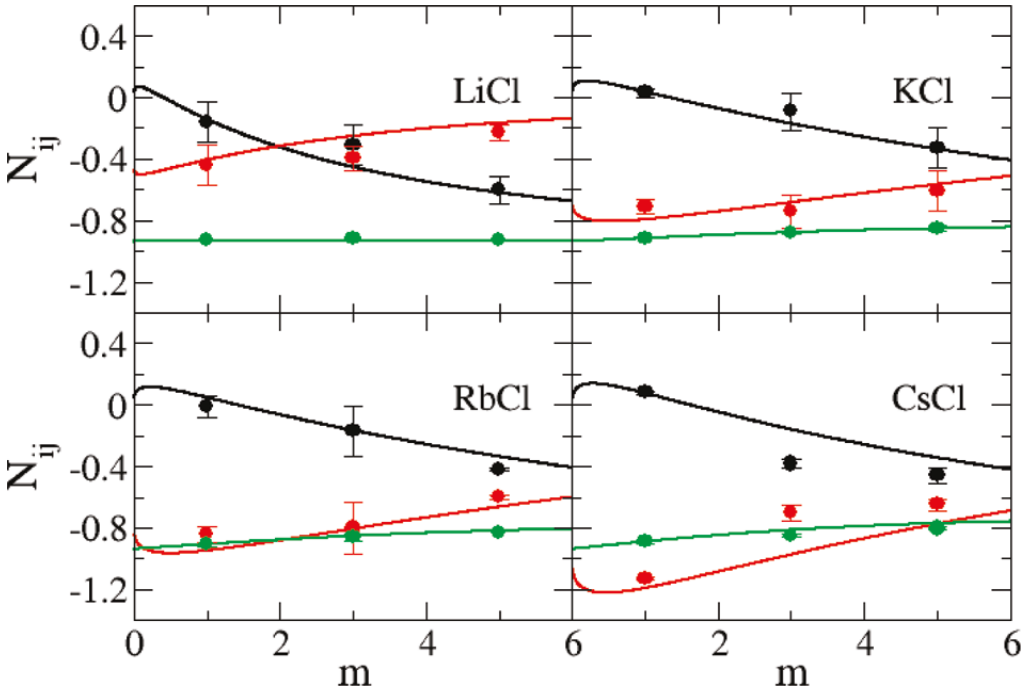
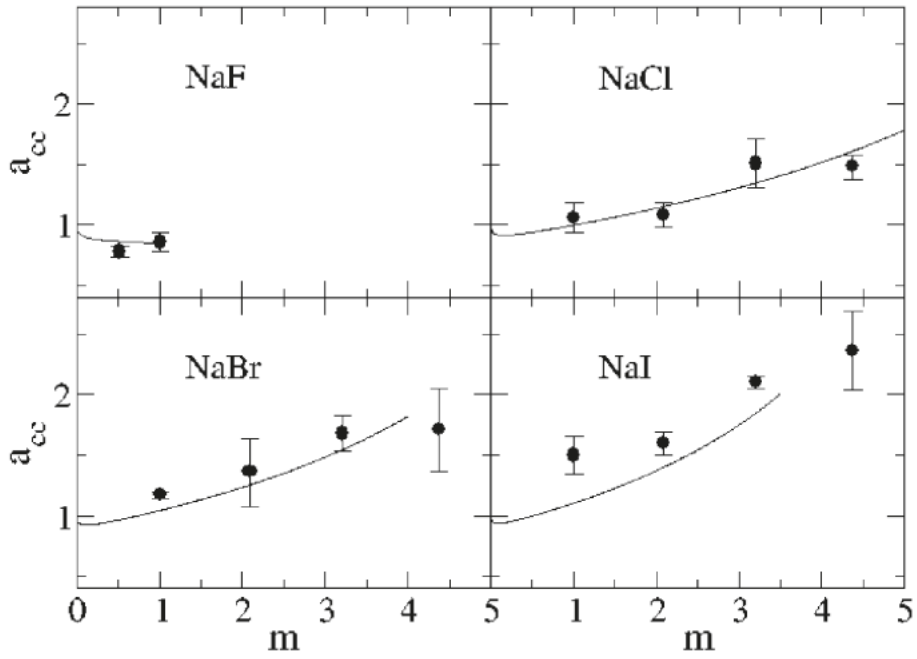


Figure A.6 Activity derivatives for sodium salts as a function of salt molality. Lines are obtained from a KB analysis of the experimental data, while symbols correspond to the results obtained with the KBFF models.



Figures A.8 and A.9 show the experimental and simulated partial molar volumes of both the water and salt as a function of the concentration. The experimental partial molar volumes of the salts generally increase monotonically, while that of water slightly decreases monotonically, as the salt concentration increases. The same trends were exhibited by the simulated values. Also, as expected, the partial molar volume of the salt increases as the size of the ions increases. The KBFF models reproduce the experimental data quantitatively except for LiCl, for which the salt partial molar volume is too large, presumably due to an overestimation of the cation size. This is also consistent with the low simulated crystal density. However, it was not possible to develop parameters using a smaller σ parameter for lithium and still reproduce the experimentally observed cation to water oxygen contact distance.Hence,we chose to correctlymodel this latter data.

Figure A.7 Activity derivatives for chloride salts as a function of salt molality. Lines are obtained from a KB analysis of the experimental data, while symbols correspond to the results obtained with the KBFF models.

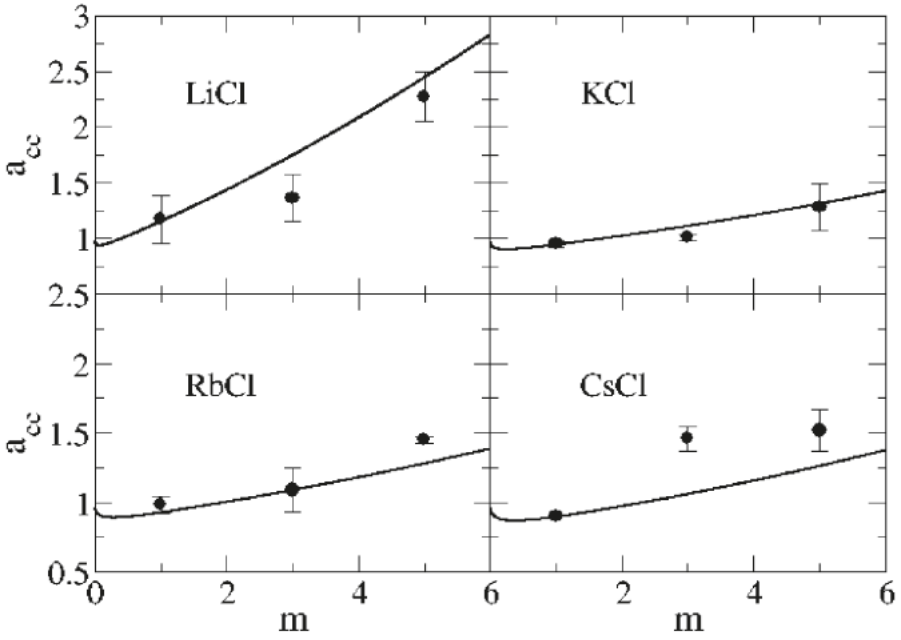


Figure A.8 Partial molar volumes (cm3/mol) for sodium salts as a function of salt molality. Lines are obtained from a KB analysis of the experimental data, while symbols correspond to the results obtained with the KBFF models. The partial molar volume of the salt is displayed in black with the partial molar volume of water displayed in red.

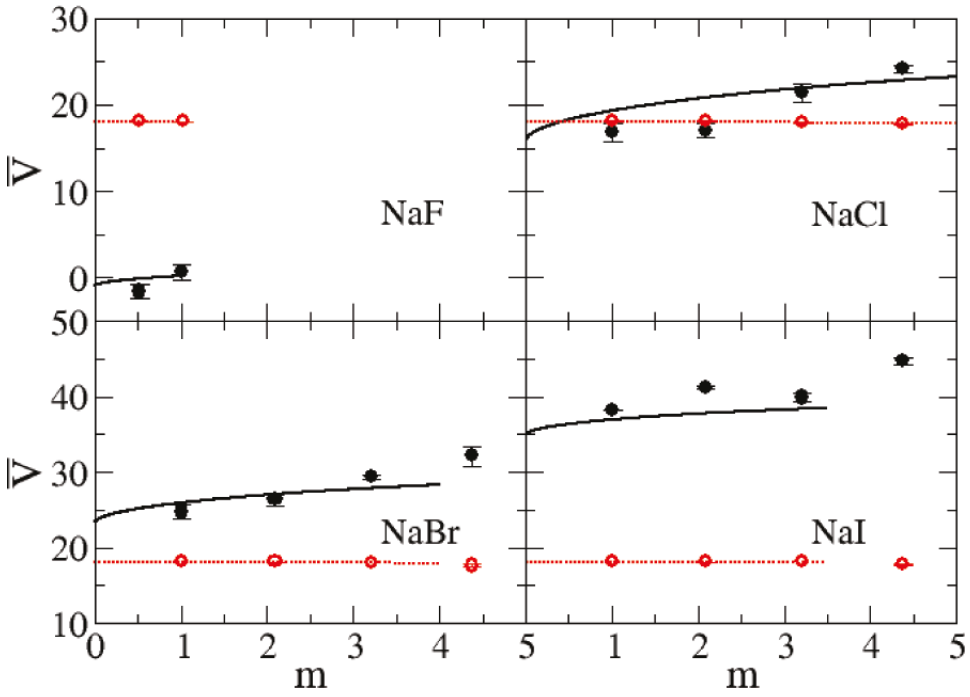
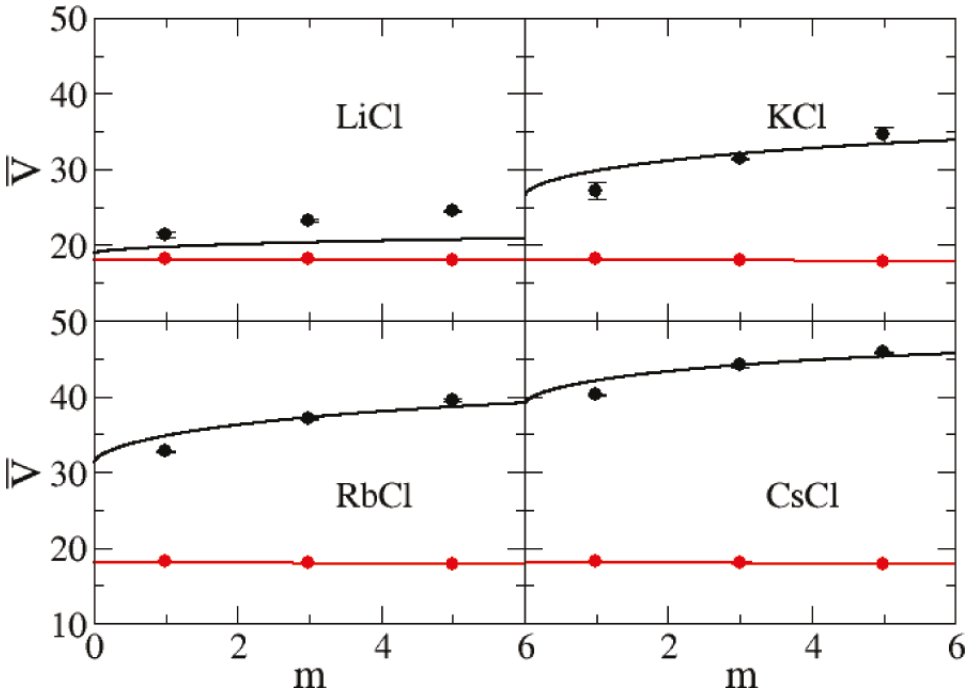


Figure A.9 Partial molar volumes (cm^3/mol) for chloride salts as a function of salt molality. Lines are obtained from a KB analysis of the experimental data, while symbols correspond to the results obtained with the KBFF models. The partial molar volume of the salt is displayed in black with the partial molar volume of water displayed in red.



The current models reproduce the excess coordination numbers, and therefore chemical potential derivatives and partial molar volumes, of a variety of salt solutions as a function of the concentration. This is the primarily goal for the KBFF models. However, it is important to test the models and their ability to reproduce other properties of salt solutions not included in the initial parametrization process, especially to see if they display significant deviations from experimental results, and to fully characterize the models in order to develop the exact range of properties for which the models will provide reliable results. The self-diffusion constants, calculated using the mean square fluctuation approach,[51] are displayed in Figures A.10 and A.11 as a function of alkali halide molality. The majority of the water, cation, and anion experimental diffusion constants all exhibit an essentially linear decrease with increasing salt molality. The notable exceptions are the diffusion constants for the chloride ion in RbCl and CsCl solutions. All the simulated diffusion constants decreased with salt concentration but typically displayed a

stronger concentration dependence compared to experimental results. The self-diffusion constants of alkali metal cations increase with size even though the mass of the ions increases, confirming that the solvation of the cation is the most important factor for the diffusion constant.[64] In contrast, the self-diffusion constants of halide ions do not display any apparent correlation with the size of the ion. Wenote, however, that it is difficult to obtain quantitative agreement with the experimental data for most solutions, as even the diffusion constant of water varies considerably between water models and can be a factor of 2 too large.[65] The agreement with experimental results can be improved somewhat by correcting for finite size effects,66 not included here, which typically result in larger (5-10%) diffusion coefficients. However, the simulated results would still appear to be more sensitive to changes in concentration compared to experimental results. It is unclear at present why this is the case. Comparison with diffusion data obtained for other models suggests the present models are reasonably competitive.[18]

Figure A.10 Diffusion constants (10^{-9} m^2/s) for sodium salts as a function of salt molality. The D_+ (black lines), D_+ (red lines), and D_w (green lines) represent the experimental diffusion constant data,[71-74] while the D_+ (black dots), D_- (red dots), and D_w (green dots) were obtained from simulations using the KBFF models.

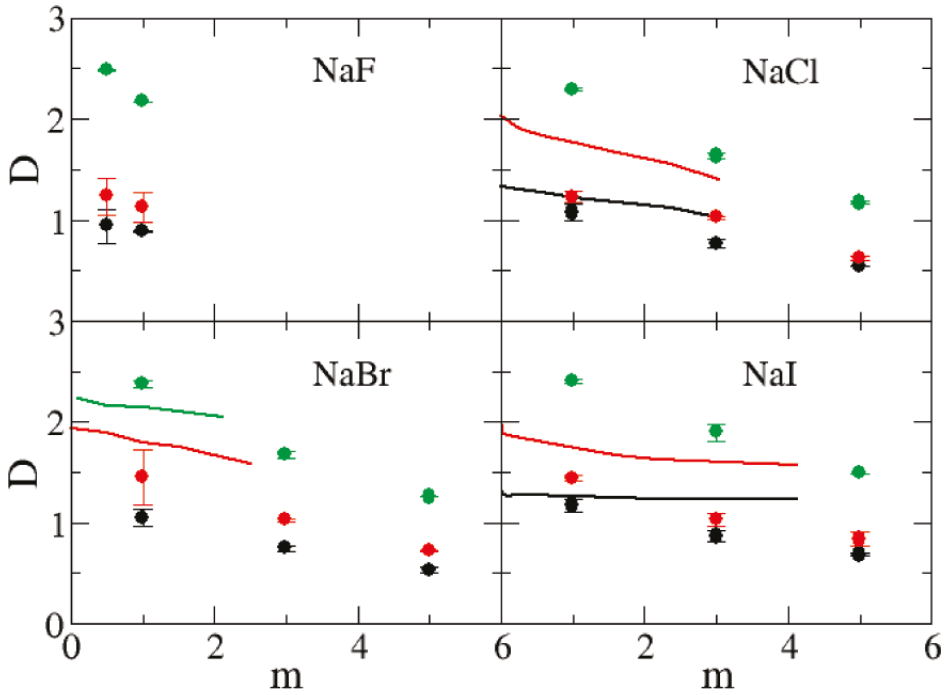
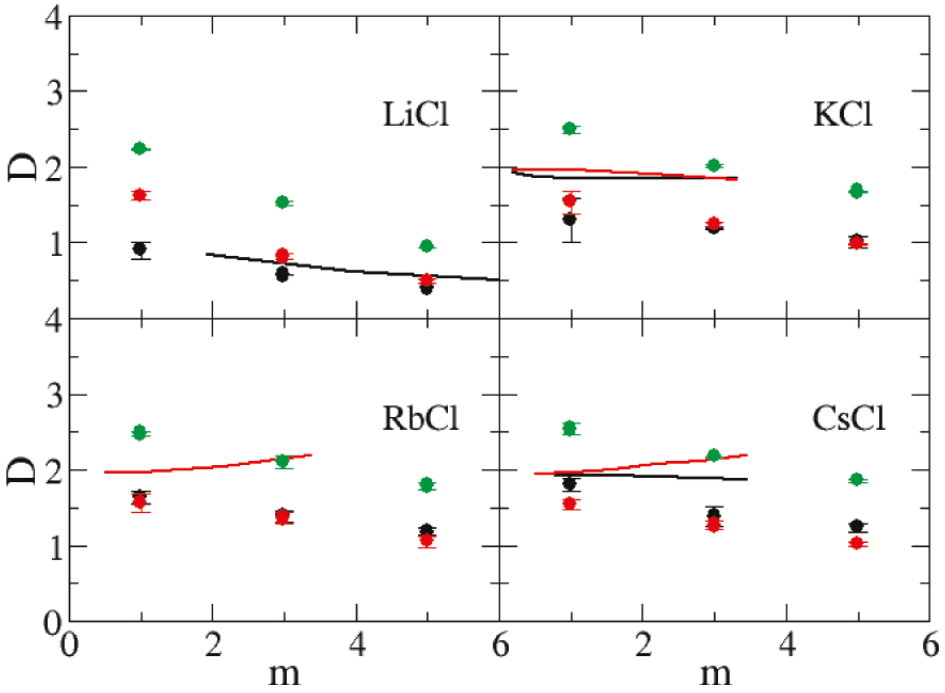


Figure A.11 Diffusion constants (10^{-9} m^2/s) for chloride salts as a function of salt molality. The D_+ (black lines), D_- (red lines), and D_w (green lines) represent the experimental diffusion constant data,[75] while the D_+ (black dots), D_- (red dots), and D_w (green dots) were obtained from simulations using the KBFF models.



The dielectric decrements (ε - ε_0) of alkali halide salts solutions, calculated from the dipole moment fluctuations,[53] are displayed in Figures A.12 and A.13. Here, ε is the relative permittivity of the solution, and ε_0 is the relative permittivity of pure water. The value of ε_0 = 63 obtained for pure water using the SPC/E model[67] is low compared to the experimental value of 78.[68] Hence, quantitative agreement for the absolute permittivities is not possible with this water model. The experimental relative permittivity for all salt solutions decreases as a function of molality, and this trend is clearly reproduced by the current models. The only exception appears to be NaF solutions at low concentrations where a small increase is observed. This increase was also reproduced in the present simulations. The KBFF models reproduce the experimental

decrement data well, with the possible exception of LiCl solutions, compared to the simulated uncertainty of ±5.

Figure A.12 Dielectric decrements ($\varepsilon - \varepsilon_0$) for a series of sodium salts as a function of salt molality. Lines were obtained from the experimental dielectric constant data,[76-78] while the symbols correspond to data obtained from simulations using the KBFF models.

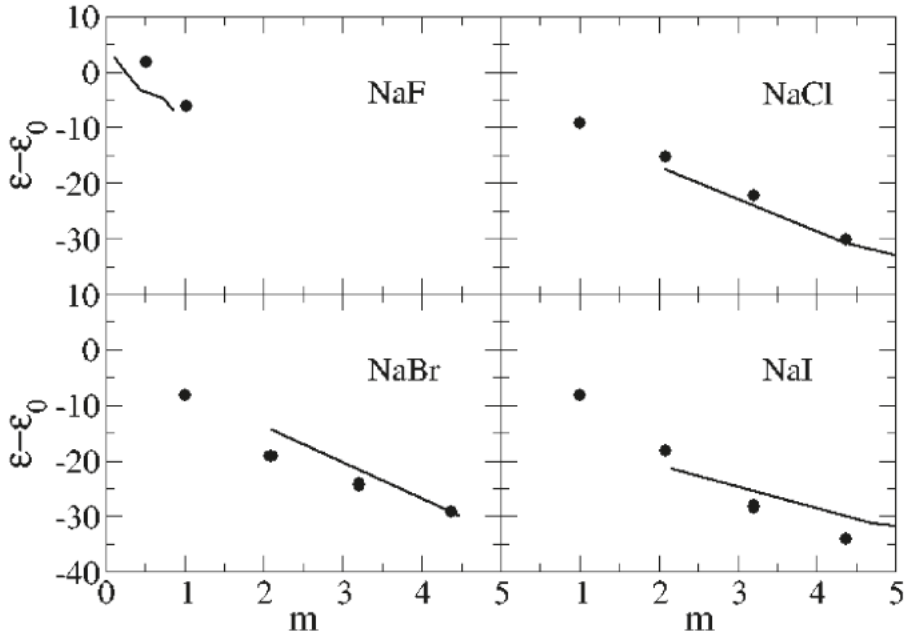
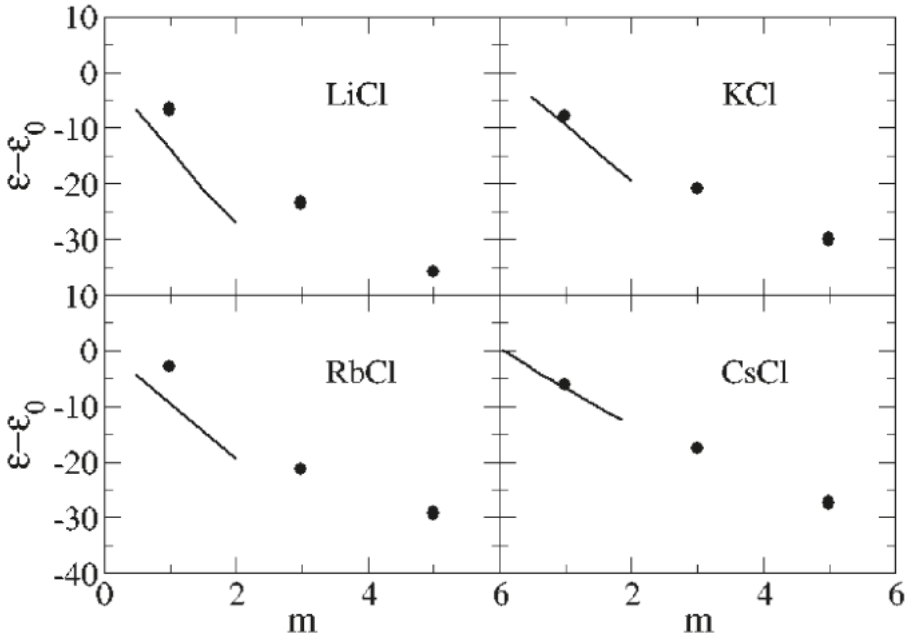


Figure A.13 Dielectric decrements ($\varepsilon - \varepsilon_0$) for a series of chloride salts as a function of salt molality. Lines were obtained from the experimental dielectric constant data,[76-78] while the symbols correspond to data obtained from simulations using the KBFF models.



The excess enthalpies of mixing for the sodium halides as a function of salt molality are displayed in Figures A.14 and A.15. The excess enthalpy of mixing for each sodium halide solution is calculated by the difference between the molar potential energy in the solution phase and in the crystal and pure water phases.[54] The data indicate that the models reproduce the experimental mixing enthalpies in a quantitative manner for NaCl, NaBr, and KCl, while the results for NaI and LiCl are somewhat too favorable. The simulated data for alkali chlorides become increasingly more unfavorable on moving from Li^+ to Rb^+ but then change sign for CsCl solutions. We presume this is due to a change in crystal structure from FCC to BCC for CsCl. It should be noted that reasonable agreement for both the free energy and enthalpy of mixing must therefore indicate good estimates for the entropy of mixing (data not shown).

Figure A.14 Excess enthalpy of mixing (kJ/mol) for sodium salts as a function of salt molality. Lines correspond to experimental data,[79] while symbols were obtained from simulations using the KBFF models.

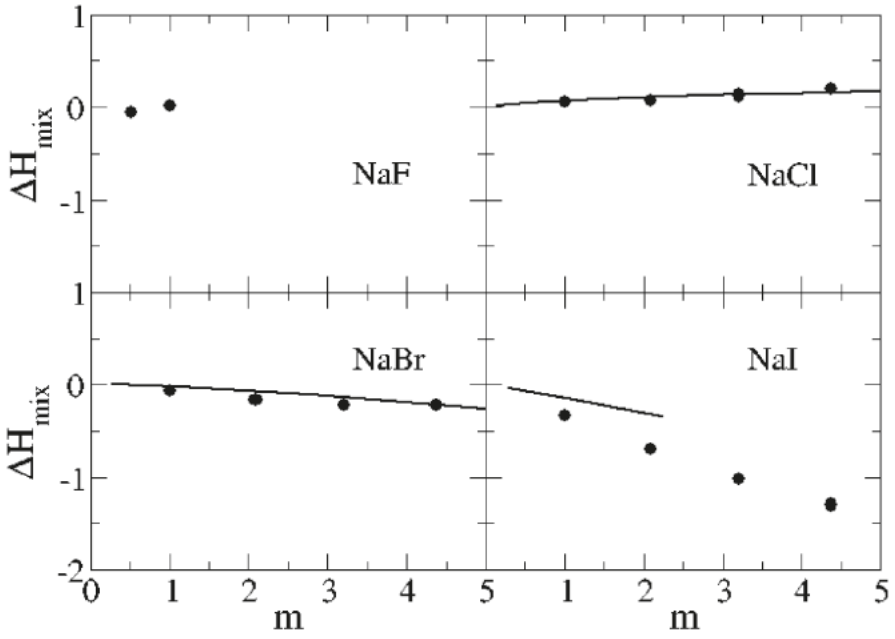
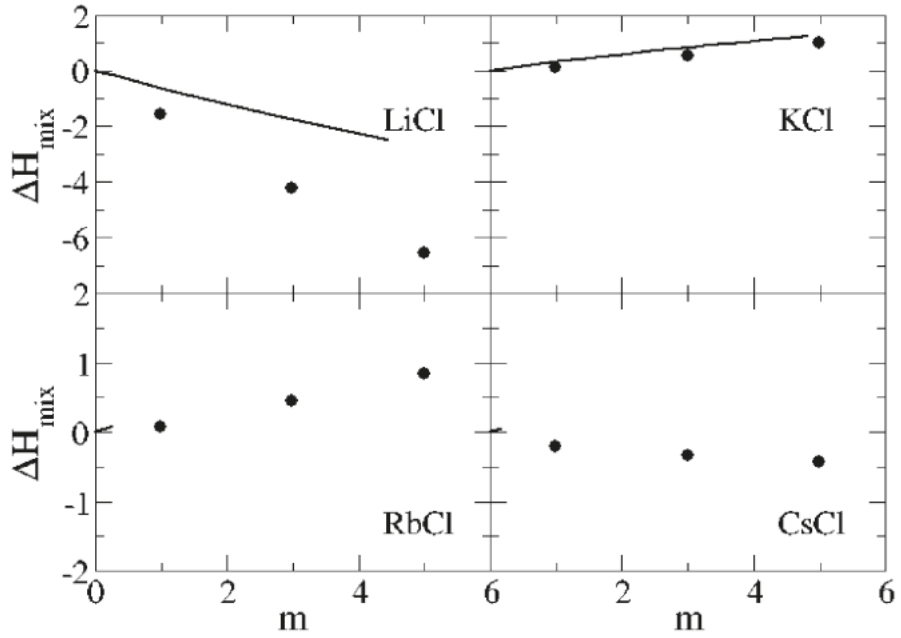


Figure A.15 Excess enthalpy of mixing (kJ/mol) for chloride salts as a function of salt molality. Lines correspond to experimental data,[79] while symbols were obtained from simulations using the KBFF models.



In the previous sections, we have developed parameters for a series of sodium halides and alkali metal chlorides by using Kirkwood-Buff theory as a guide. In order to demonstrate the transferability of the parameters to a variety of alkali halides, we have used the same ion parameters to study two other systems, aqueous KI and aqueous CsBr, which were not included in the previous parametrization and for which there are no longer any free parameters. The results are presented in Figures A.16-A.18 and clearly suggest that, to a high degree of accuracy, the parameters developed here for the sodium and chloride salts are transferable to other alkali halide salts.

Figure A.16 Excess coordination numbers as a function of salt molality (top). The N_{cc} (black lines), N_{cw} (red lines), and N_{ww} (green lines) are obtained from a KB analysis of the experimental data. The N_{cc} (black dots), N_{cw} (red dots), and N_{ww} (green dots) are obtained from simulations. Activity derivatives as a function of salt molality (bottom): Lines are obtained from a KB analysis of the experimental data, while symbols correspond to results obtained using the KBFF models.

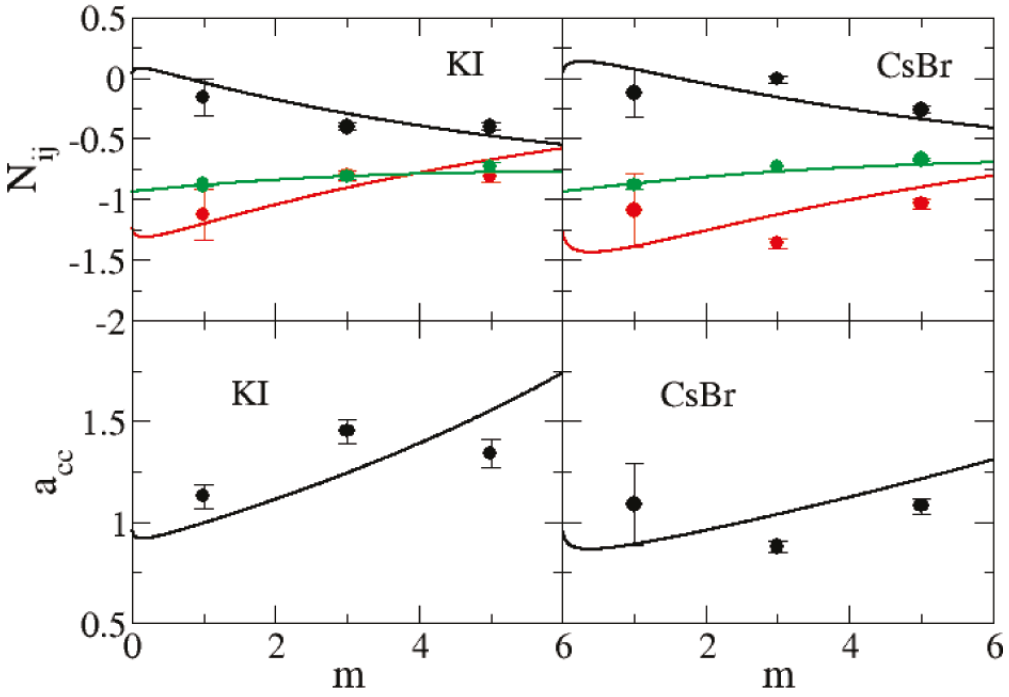


Figure A.17 Partial molar volumes (cm^3/mol) as a function of salt molality (top). Lines are obtained from a KB analysis of the experimental data, while symbols correspond to results obtained using the KBFF models. The black lines and symbols represent the partial molar volume of the salt, while red lines and symbols indicate partial molar volume of water. Diffusion constants (10^{-9} m^2/s) as a function of salt molality (bottom): The D_+ (black lines), D_- (red lines), and D_w (green lines) are obtained from experimental diffusion constant data,[80] while the D_+ (black •), D_- (red ○), and D_w (green x) were obtained from simulations performed using the KBFF models.

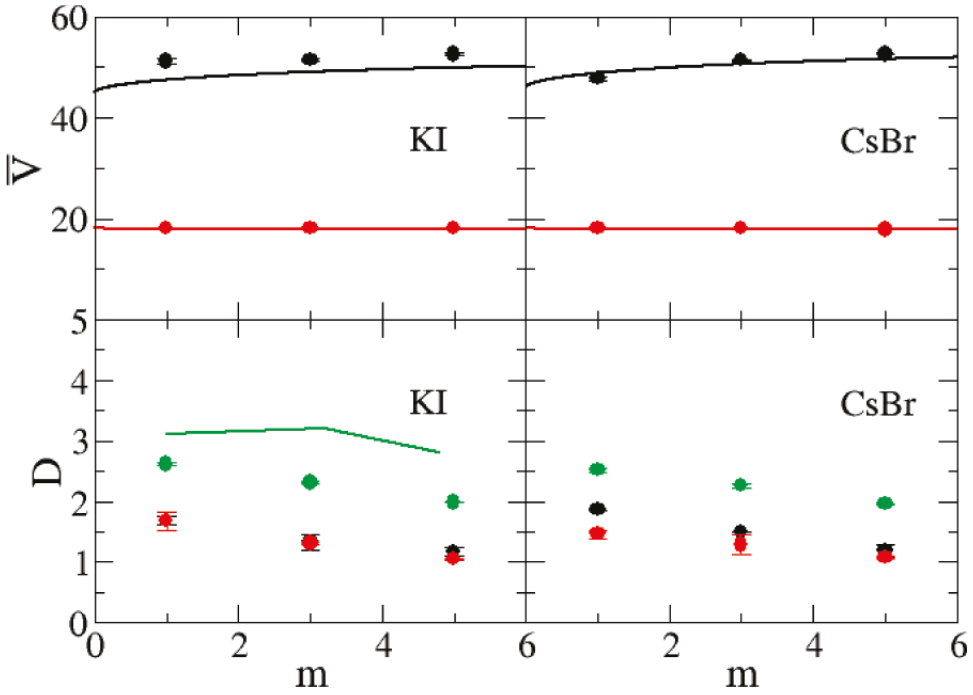
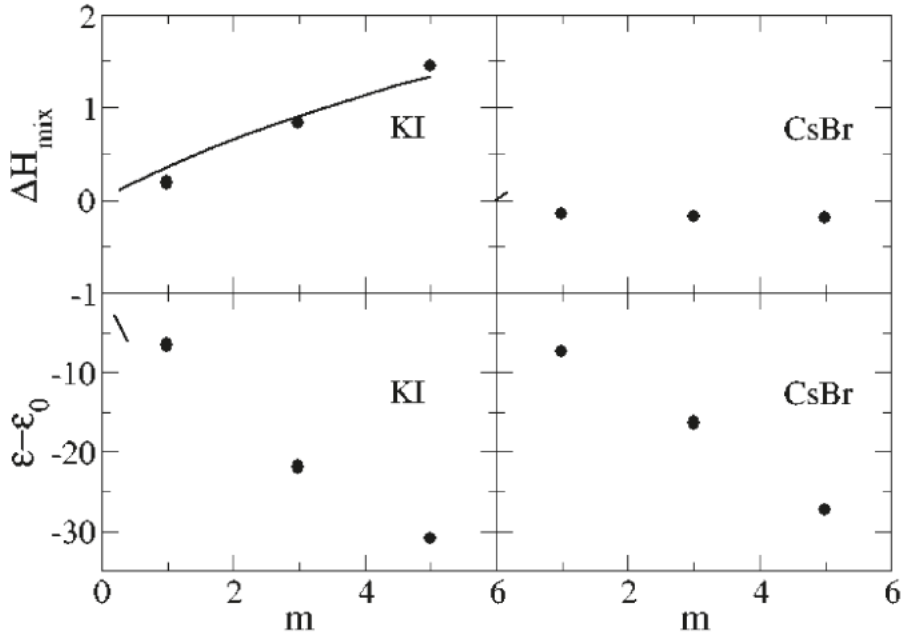


Figure A.18 Excess enthalpy of mixing (kJ/mol) as a function of salt molality (top) and dielectric decrements as a function of salt molality (bottom). Lines correspond to the experimental data,[79] while symbols were obtained from simulations using the KBFF models.



A.5 Conclusions

A series of models for aqueous alkali halide solutions have been developed by attempting to reproduce the experimentally derived Kirkwood-Buff integrals using molecular dynamics simulation. A major advantage of this type of approach is the ability to provide insight into salt activities in a computationally efficient manner and to ensure a reasonably accurate balance between solute-solute (N_{cc}) and solute-solvent (N_{cw}) distributions and, by inference, their interactions. Other physical and thermodynamic properties such as ion diffusion constants, relative permittivity, density, and heat of mixing have also been reasonably well reproduced. In addition, by examining the results obtained for aqueous KI and CsBr solutions, it has been clearly demonstrated that the parameters developed for sodium and chloride salts are transferable to other alkali halide salts. Unfortunately, not all the models provide good agreement for all the experimental data. To some degree, this is expected when using such simple models. The major

issues involved the most highly polarizing ions (Li^+ an F^-), while the diffusion constant data also provided only modest agreement with experimental results. Hence, care should be taken when using the current models for these types of applications. The models are specifically designed to be used with the SPC/E water model, although, according to previous studies,[25,69] other simple point charge models should provide similar results. The recent models contribute to a consistent set of parameters that can eventually be used to study salt effects on peptides and proteins.

The solutions studied in this work include a variety of polarizable and polarizing anions and cations over a range of compositions. It is encouraging that one can reproduce much of the experimental data with the simple nonpolarizable models used here. However, to achieve this goal, it was necessary to break the standard combination rules when determining the cation-water interactions. The modified ε parameters actually lead to an increase in the cation-water interaction and can be thought of, to some degree, as a crude approach to incorporate polarization effects, which undoubtedly play a significant role in these solutions.

The present models provide an alternative to other recent ion force fields developed using more traditional approaches—such as the free energy of hydration.We have argued that the use of the experimental KBIs provides a rigorous test of force field accuracy and thereby provides ideal target data for the parametrization.[31] Furthermore, this can be achieved without a significant sacrifice in agreement with other solution properties. Whether the current models are substantially better than other, more traditional, models remains to be seen. This issue requires a more thorough and comprehensive study than is feasible here. The present models should be viewed as providing a reasonable balance between solute-solute, solute-solvent, and solvent-solvent interactions, as inferred by their resulting distributions, and are therefore suitable for studies of solute activities and cosolvent interactions with biomolecules.[30,63] Of course, one should always test that any potential model reasonably reproduces any specific properties of interest before use.

A.6 Supporting Information

Supporting Information is provided which includes tables containing a summary of all the simulations performed in this study, first shell coordination numbers, fitting constants for both the experimental and simulated activity data (eq A.4), and a comparison of the present lattice energies with experimental and other simulation data.

Table A.4 Summary of the MD simulations of alkali halide water mixtures: All simulations were performed at 300 K and 1atm in the *NpT* ensemble. Symbols are N_w, number of water molecules; N_s (= $N_+ = N_- = 1/2N_c$), number of alkali-halide pairs; V, average simulation volume; ms, salt molality; C_s, salt molarity; ρ, mass density; E_{pot}, average total potential energy per molecule ($N_s + N_w$); and T_{sim}, total simulation time.

	N_s	N_w	m_s (mol/kg)	V (nm^3)	C_s (mol/l)	ρ (g/cm^3)	E_{pot} (kJ/mol)	T_{sim} (ns)
H2O	0	2170	0.00	65.265	0.00	0.995	-46.45	2
NaF	20	2150	0.52	64.531	0.52	1.018	-54.94	6
	38	2079	1.01	64.519	1.03	1.040	-63.05	6
NaCl	38	2079	1.01	63.595	0.99	1.036	-60.07	6
	77	2048	2.09	63.829	2.00	1.077	-73.99	4
	115	1987	3.21	63.354	3.01	1.114	-88.00	4
	154	1950	4.38	63.783	4.01	1.149	-102.02	4
NaBr	38	2079	1.01	64.089	0.98	1.072	-59.62	5
	77	2048	2.09	64.810	1.97	1.148	-73.05	5
	115	1987	3.21	64.730	2.95	1.222	-86.59	9
	154	1950	4.38	64.584	3.90	1.291	-100.08	5
	231	1730	7.41	63.426	6.05	1.438	-132.53	9
	308	1600	10.69	64.350	7.95	1.562	-163.90	9
NaI	38	2079	1.01	65.051	0.97	1.101	-58.86	5
	77	2048	2.09	66.683	1.92	1.206	-71.52	5
	115	1987	3.21	67.458	2.83	1.305	-84.27	5
	154	1950	4.38	69.151	3.70	1.398	-96.97	5
	231	1730	7.41	68.599	5.59	1.593	-127.35	5
	308	1600	10.69	71.253	7.18	1.748	-156.79	5
LiCl	127	7065	1.00	216.903	0.97	1.016	-62.37	6
	367	6796	3.00	217.842	2.80	1.052	-92.29	6
	589	6541	5.00	219.083	4.47	1.082	-120.13	6
KCl	126	7002	1.00	216.178	0.97	1.041	-58.36	6

	357	6603	3.00	215.924	2.75	1.120	-80.74	6
	561	6228	5.00	216.204	4.31	1.183	-101.56	6
RbCl	125	6963	1.00	216.055	0.96	1.080	-57.79	6
	352	6512	3.00	215.995	2.71	1.229	-79.14	6
	549	6093	5.00	216.324	4.22	1.352	-99.03	6
CsCl	125	6915	1.00	216.032	0.96	1.119	-57.38	6
	345	6385	3.00	215.737	2.66	1.333	-77.74	6
	533	5920	5.00	216.025	4.10	1.510	-96.76	6
KI	124	6880	1.00	217.183	0.95	1.105	-57.18	6
	340	6300	3.00	218.381	2.59	1.292	-77.20	6
	522	5796	5.00	219.331	3.95	1.447	-95.97	6
CsBr	124	6870	1.00	216.248	0.95	1.153	-56.93	6
	339	6275	3.00	216.377	2.60	1.422	-76.47	6
	519	5761	5.00	216.598	3.98	1.640	-94.77	6

Table A.5 First shell coordination numbers (n_{ij}^{1st}) and second shell coordination numbers (n_{ij}^{2nd}) as a function of concentration (m) alkali halide aqueous solutions. $R_{max}^{1st}/R_{min}^{1st}$ and $R_{max}^{2nd}/R_{min}^{2nd}$ are the distances (nm) to the first and the second maximum/minimum of the radial distribution functions. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

		m	+/-	+/o	-/o	o/o
NaF	$R_{max}^{1st}/$		0.25	0.23	0.30	0.27
	R_{min}^{1st}		0.326	0.310	0.364	0.342
	n_{ij}^{1st}	0.5	0.06	5.54	7.01	4.80
		1.0	0.03	5.48	7.04	4.75
	R_{max}^{2nd}		0.48	0.44	0.48	0.45
	R_{min}^{2nd}		0.562	0.540	0.592	0.564
	n_{ij}^{2nd}	0.5	0.39	17.51	21.90	19.29
		1.0	0.80	17.38	21.85	19.25
NaCl	$R_{max}^{1st}/$		0.27	0.23	0.32	0.28
	R_{min}^{1st}		0.355	0.315	0.405	0.345
	n_{ij}^{1st}	0.99	0.09	5.52	8.11	5.12
		2.00	0.20	5.39	8.22	5.11
		3.01	0.42	5.11	8.38	5.06
		4.01	0.57	4.93	8.45	5.00
NaBr	$R_{max}^{1st}/$		0.28	0.23	0.33	0.28
	R_{min}^{1st}		0.365	0.315	0.415	0.405
	n_{ij}^{1st}	0.98	0.10	5.49	7.54	5.09
		1.97	0.22	5.34	7.62	6.57
		2.95	0.35	5.17	8.11	7.06
		3.90	0.50	4.97	8.61	8.11
		6.05	0.96	4.33	9.66	8.72
		7.95	1.45	3.67	10.31	8.03
NaI	$R_{max}^{1st}/$		0.29	0.23	0.35	0.28
	R_{min}^{1st}		0.375	0.315	0.425	0.395

	n_{ij}^{1st}	0.97	0.07	5.48	7.91	5.04
		1.92	0.16	5.40	7.98	6.42
		2.83	0.26	5.24	8.04	6.83
		3.70	0.39	5.05	8.59	7.18
		5.59	0.84	4.37	9.60	7.56
		7.18	1.31	3.66	10.80	6.79
LiCl	$R_{max}^{1st}/R_{min}^{1st}$		0.23 / 0.315	0.19 / 0.265	0.32 / 0.398	0.28 / 0.360
	n_{ij}^{1st}	1	0.04	3.96	7.67	5.51
		3	0.12	3.88	7.80	5.40
		5	0.22	3.78	7.91	5.25
	$R_{max}^{2nd}/R_{min}^{2nd}$		0.46 / 0.546	0.41 / 0.502	0.50 / 0.626	0.45 / 0.566
	n_{ij}^{2nd}	1	0.61	14.16	25.07	18.59
		3	1.48	13.29	24.32	17.29
		5	2.26	12.46	23.36	16.27
KCl	$R_{max}^{1st}/R_{min}^{1st}$		0.31 / 0.389	0.26 / 0.342	0.32 / 0.384	0.28 / 0.334
	n_{ij}^{1st}	1	0.20	6.11	7.03	4.32
		3	0.54	5.77	6.79	4.03
		5	0.89	5.39	6.48	3.74
RbCl	$R_{max}^{1st}/R_{min}^{1st}$		0.32 / 0.404	0.28 / 0.361	0.32 / 0.385	0.27 / 0.332
	n_{ij}^{1st}	1	0.24	6.75	7.03	4.23
		3	0.60	6.35	6.72	3.89
		5	0.95	5.95	6.32	3.57
CsCl	$R_{max}^{1st}/R_{min}^{1st}$		0.34 / 0.428	0.29 / 0.376	0.32 / 0.385	0.27 / 0.334
	n_{ij}^{1st}	1	0.34	7.15	6.91	4.28
		3	0.78	6.63	6.48	3.92

		5	1.18	6.15	6.06	3.59
KI	$R_{max}^{1st}/$ R_{min}^{1st}		0.33 0.418	0.26 0.342	0.34 0.412	0.27 0.332
	n_{ij}^{1st}	1	0.24	6.02	7.34	4.20
		3	0.68	5.47	7.00	3.83
		5	1.10	4.96	6.61	3.51
CsBr	$R_{max}^{1st}/$ R_{min}^{1st}		0.34 0.438	0.29 0.380	0.33 0.394	0.27 0.332
	n_{ij}^{1st}	1	0.39	7.18	6.90	4.19
		3	0.93	6.47	6.34	3.81
		5	1.38	5.88	5.85	3.49

Table A.6 Fitting constants for Equation 4 determined by fitting experimental alkali halide activity coefficients and the corresponding simulated activity derivatives.

		a_1	a_2	a_3	a_4	m_s^{max}
NaF	Exp	1.2759	-0.0410	0	0	1.0
	KBFF	1.2759	0.1757	0	0	1.0
NaCl	Exp	1.3360	0.0810	0	0	6.0
	KBFF	1.3360	0.0827	0	0	5.0
NaBr	Exp	1.4360	0.1041	0	0	4.0
	KBFF	1.4360	0	0.3342	0.0263	5.0
NaI	Exp	1.6350	0.1300	0	0	3.5
	KBFF	1.6350	0	0.4308	0.0133	5.0
LiCl	Exp	1.4644	0	0.2110	0.0102	6.0
	KBFF	1.4644	0	-0.2430	0.0086	5.0
KCl	Exp	1.1720	0.0360	0	0	5.0
	KBFF	1.1720	0.0260	0	0	5.0
RbCl	Exp	1.0643	0.0311	0	0	5.0
	KBFF	1.0643	0.0482	0	0	5.0
CsCl	Exp	0.8402	0.0311	0	0	6.0
	KBFF	0.8402	0.0398	0	0	5.0
KI	Exp	1.4259	0.0480	0	0	4.5
	KBFF					N/A
CsBr	Exp	0.8453	0.0216	0	0	5.0
	KBFF					N/A

Experimental data was fitted to Equation A.4

$$ln\gamma_{\pm} = \frac{-1.178\sqrt{m_s}}{1 + a_1\sqrt{m_s}} - ln(1 - a_2 m_s) + a_3 m_s + a_4 m_4^2$$

(5.1)

Simulated data was fitted using the following activity coefficient derivative,

$$\left(\frac{\partial ln\gamma_{\pm}}{\partial lnm_s}\right)_{T,P} = \frac{-N_{22}^{+}}{1+N_{22}^{+}} \qquad N_{22}^{+} = N_{22} + m_2(1 + N_{11} - 2N_{21})$$

(5.2)

which can be then be expressed in terms of the derivative of Equation A.4. The fits should be considered approximate due to the relatively low number of points used (typically three). The simulated data for KI and CsBr could not be fitted accurately using the above relationships.

Table A.7 Lattice energies (kJ/mol) for alkali halide crystals.

	Exp	Hagler FF	KBFF
NaF	895.4	937.2	1217.7
NaCl	765.7	786.6	808.2
NaBr	728.0	744.8	776.1
NaI	682.0	698.7	751.0
LiCl	832.6	861.9	1178.0
KCl	694.5	711.3	725.3
RbCl	665.3	686.2	692.7
CsCl	648.5	644.3	650.1
KI	627.6	636.0	663.2
CsBr	627.6	615.0	628.8

Exp: Experimental data taken from (Tosi, M. P.; Fumi, F. G. J Phys Chem Solids 1964, 25, 45)
Hagler FF: Hagler force field data taken from (Peng, Z. W.; Ewig, C. S.; Hwang, M. J.; Waldman, M.; Hagler, A. T. J Phys Chem A 1997, 101, 7243)

Figure A.19 Radial distribution functions of NaF obtained from the 0.5 m (black lines) and 1 m (red lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

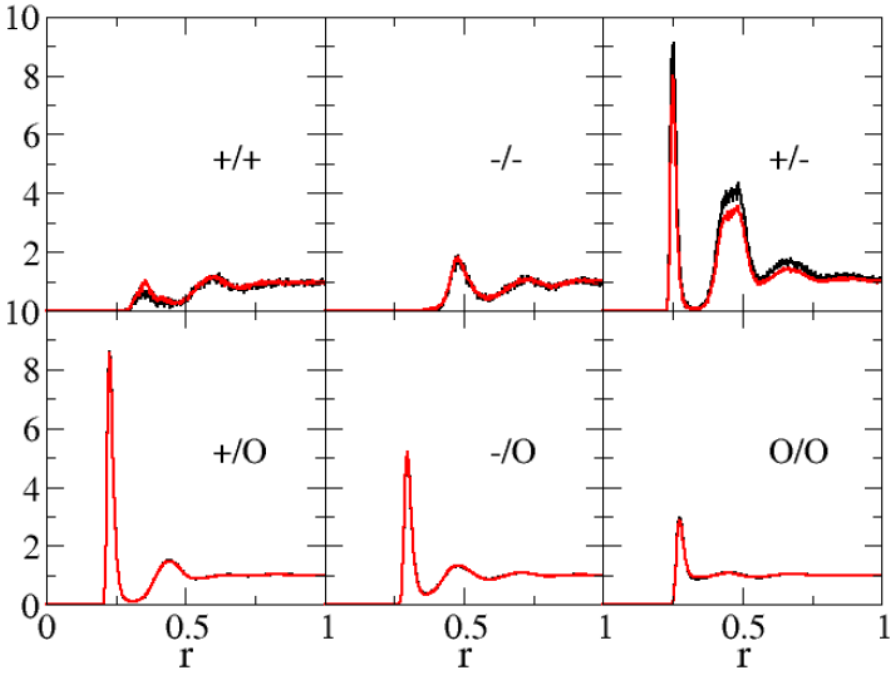


Figure A.20 Radial distribution functions of NaCl obtained from the 0.99 m (black lines) and 4.01 m (red lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

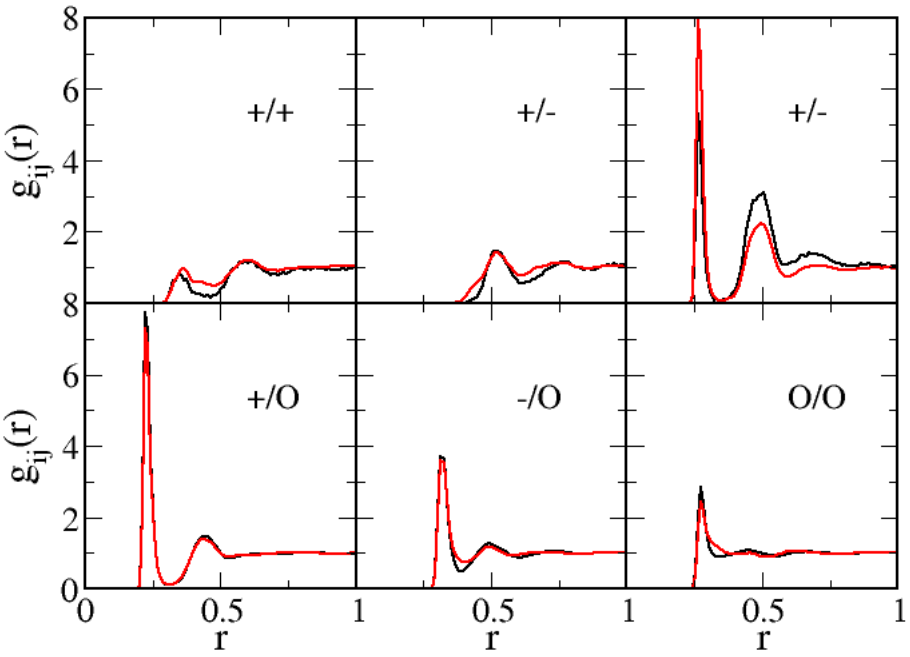


Figure A.21 Radial distribution functions of NaBr obtained from the 0.98 m (black lines), 2.95 m (red lines), and 6.05 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

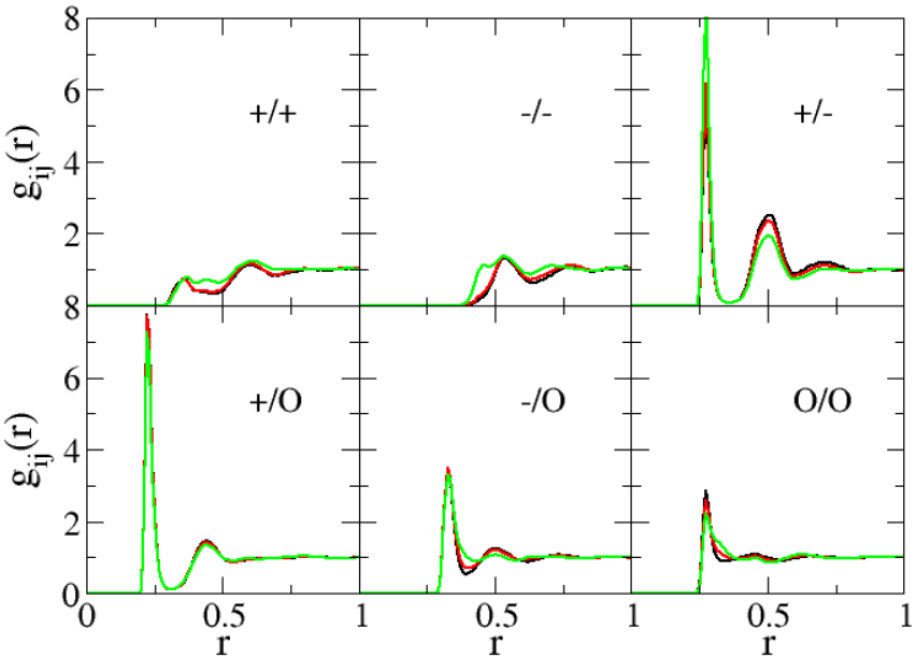


Figure A.22 Radial distribution functions of NaI obtained from the 0.97 m (black lines), 2.83 m (red lines), and 5.59 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

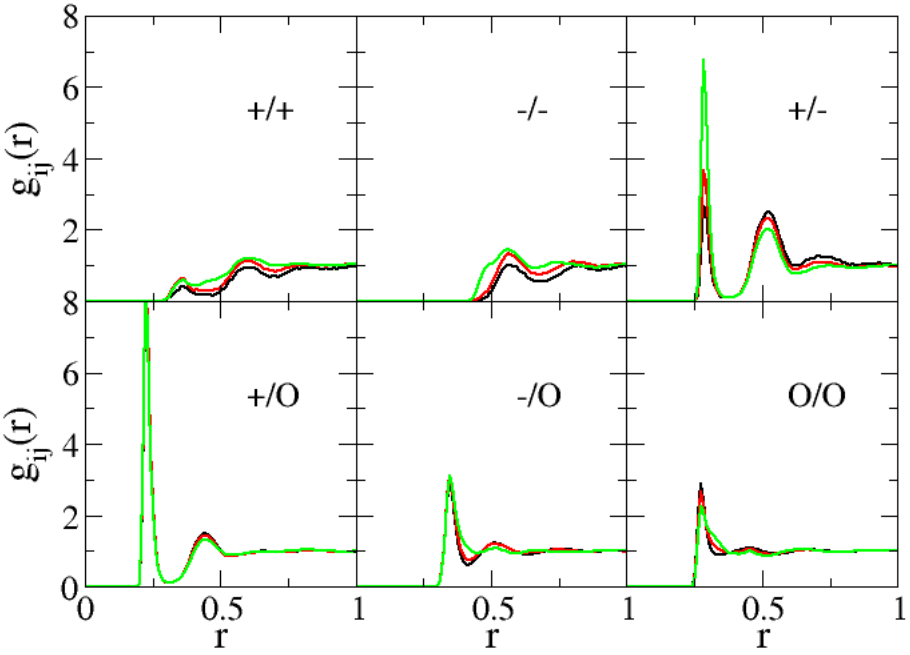


Figure A.23 Radial distribution functions of LiCl obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

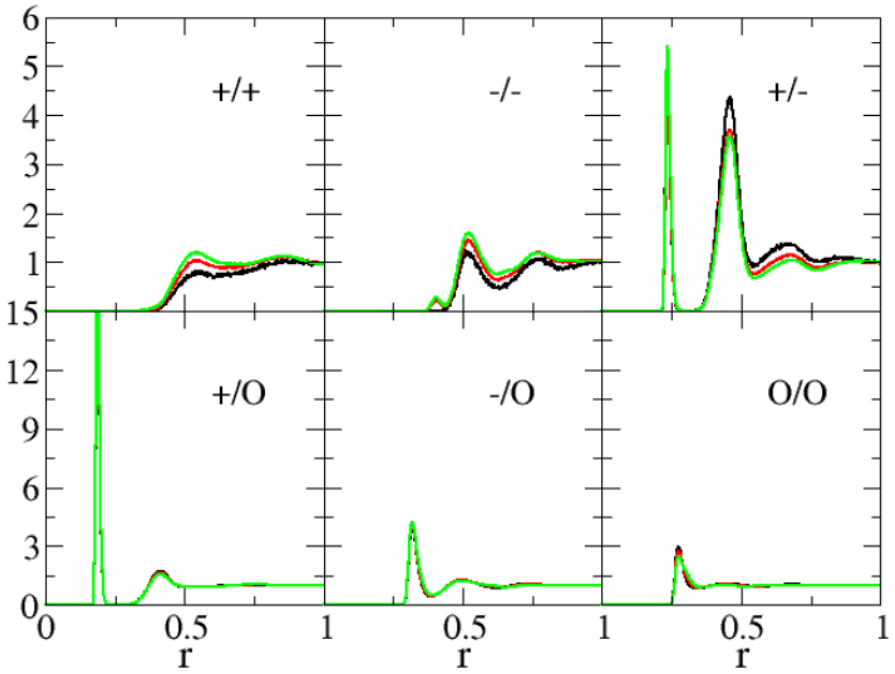


Figure A.24 Radial distribution functions of KCl obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

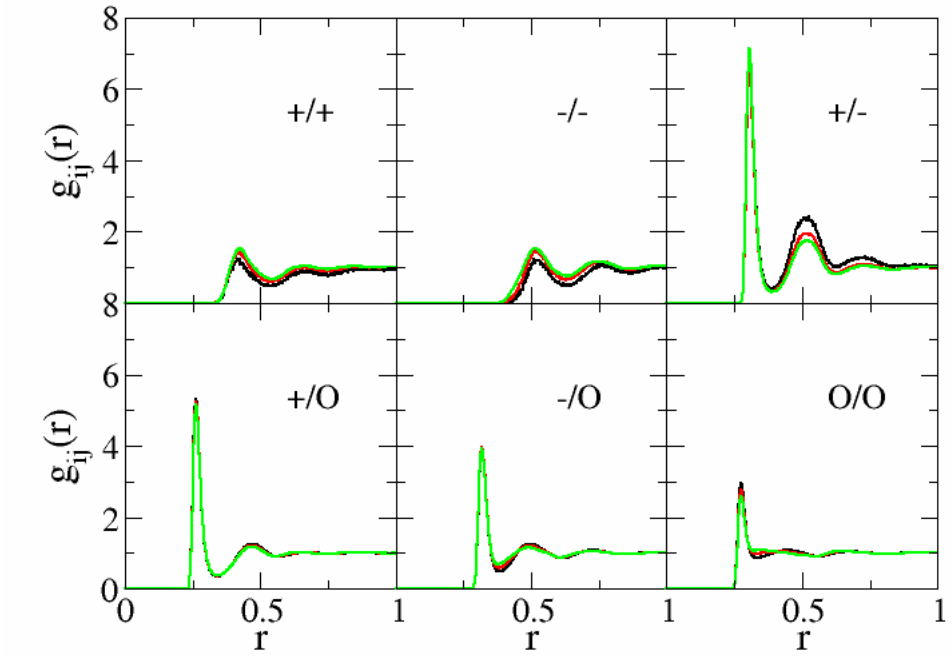


Figure A.25 Radial distribution functions of RbCl obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

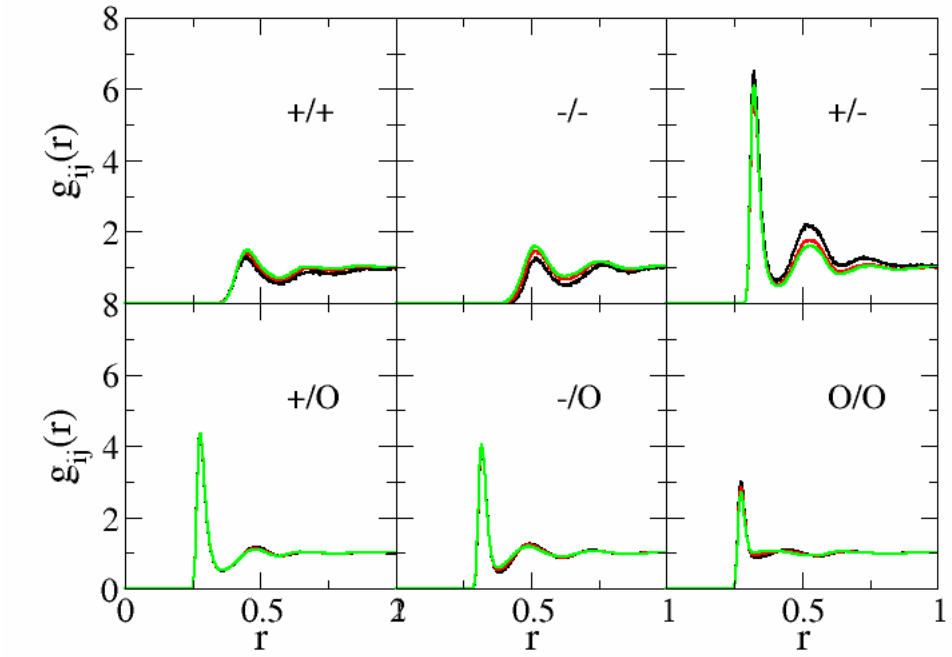


Figure A.26 Radial distribution functions of CsCl obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

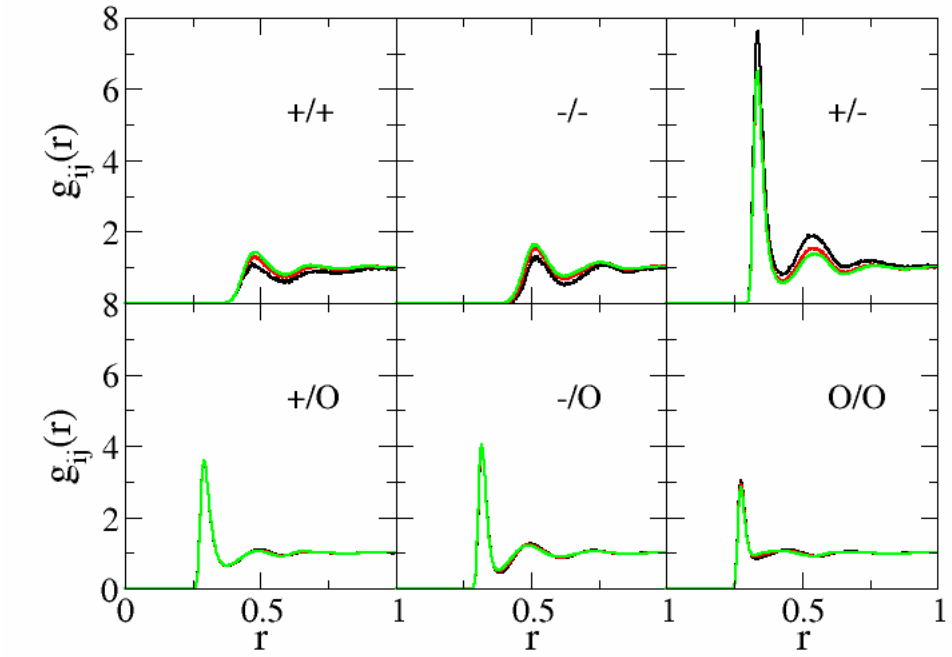


Figure A.27 Radial distribution functions of KI obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.

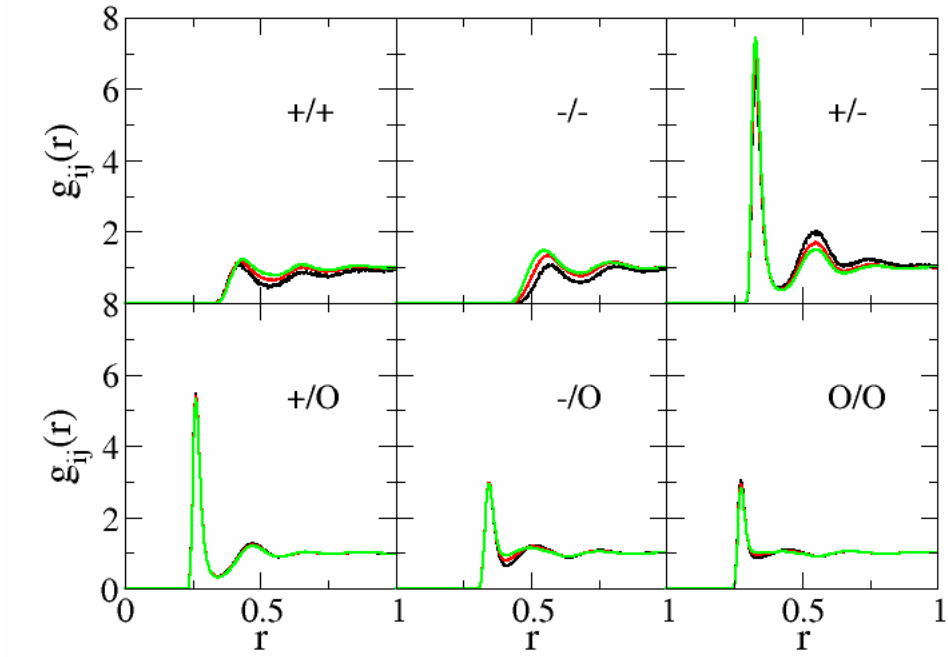
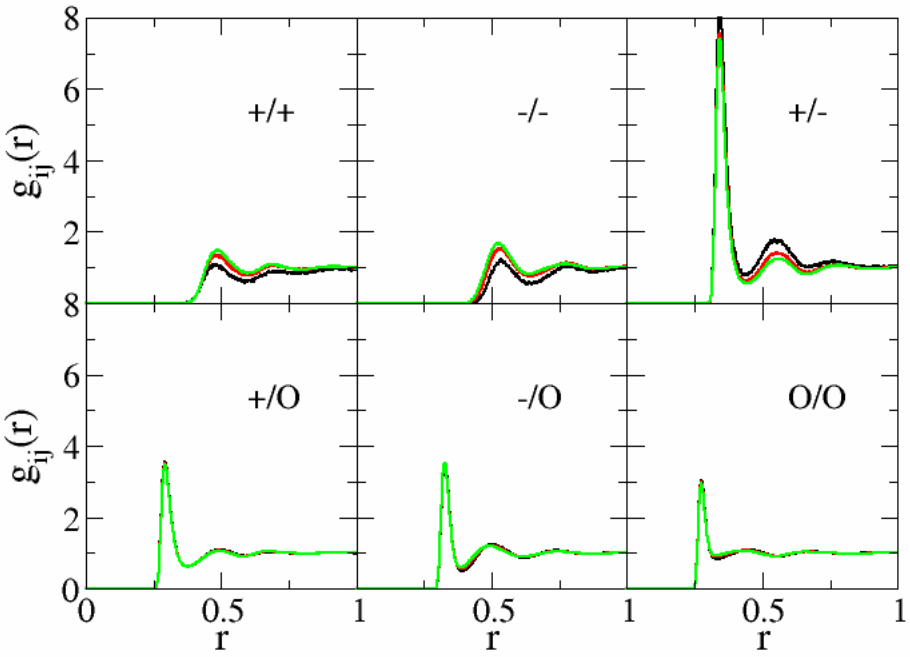


Figure A.28 Radial distribution functions of CsBr obtained from the 1 m (black lines), 3 m (red lines), and 5 m (green lines) simulations. Cations, anions, and the water oxygen are denoted by the symbols +, -, and o, respectively.



References

(1) Mclaughlin, S. *Annual Review Biophysics and Biology* **1989**, 18, 113–136.

(2) Anderson, C. F.; Record, M. T. *Annual Review Physical Chemistry* **1995**, 46, 657–700.

(3) Baldwin, R. L. *Biophysics Journal* **1996**, 71, 2056–2063.

(4) Joung, I. S.; Cheatham, T. E. *Journal of Physical Chemistry B* **2008**, 112, 9020–9041.

(5) Aqvist, J. *Journal of Physical Chemistry* **1990**, 94, 8021–8024.

(6) Peng, Z. W.; Ewig, C. S.; Hwang, M. J.; Waldman, M.; Hagler, A. T. *Journal of Physical Chemistry A* **1997**, 101, 7243–7252.

(7) Rasaiah, J. C. *Journal of Chemical Physics* **1970**, 52, 704–715.

(8) Lee, S. H.; Rasaiah, J. C. *Journal of Chemical Physics* **1994**, 101, 6964–6974.

(9) Du, H.; Rasaiah, J. C.; Miller, J. D. *Journal of Physical Chemistry B* **2007**, 111, 209–217.

(10) Lamoureux, G.; Roux, B. *Journal of Physical Chemistry B* **2006**, 110, 3308–3322.

(11) Jensen, K. P.; Jorgensen, W. L. *Journal of Chemical Theory and Computation* **2006**, 2, 1499–1509.

(12) Chen, A. A.; Pappu, R. V. *Journal of Physical Chemitry B* **2007**, 111, 11884–11887.

(13) Auffinger, P.; Cheatham, T. E.; Vaiana, A. C. *Journal of Chemical Theory and Computation* **2007**, 3, 1851–1859.

(14) Smith, D. E.; Dang, L. X. *Journal of Chemical Physics* **1994**, 100, 3757–3766.

(15) Dang, L. X. *Journal of Chemical Physics* **1992**, 96, 6970–6977.

(16) Dang, L. X.; Garrett, B. C. *Journal of Chemical Physics* **1993**, 99, 2972–2977.

(17) Horinek, D.; Mamatkulov, S. I.; Netz, R. R. *Journal of Chemical Physics* **2009**, 130, 124507.

(18) Joung, I. S.; Cheatham, T. E. *Journal of Physical Chemistry B* **2009**, 113, 13279–13290.

(19) Bentenitis, N.; Cox, N. R.; Smith, P. E. *Journal of Physical Chemistry B* **2009**, 113, 12306–12315.

(20) Kang, M.; Smith, P. E. *Journal of Computational Chemistry* **2006**, 27, 1477–1485.

(21) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2005**, 109, 15080–15086.

(22) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2004**, 121, 2180–2186.

(23) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, 118, 10663–10670.

(24) Weerasinghe, S.; Smith, P. E. *Journal of Physical Chemistry B* **2003**, 107, 3891–3898.

(25) Weerasinghe, S.; Smith, P. E. *Journal of Chemical Physics* **2003**, 119, 11342–11349.

(26) Chitra, R.; Smith, P. E. *Journal of Physical Chemistry B* **2002**, 106, 1491–1500.

(27) Smith, P. E. *Journal of Chemical Physics* **2008**, 129, 124509.

(28) Ben-Naim, A. *Journal of Chemical Physics* **1977**, 67, 4884–4890.

(29) Ben-Naim, A. *Statistical Thermodynamics for Chemists and Biochemists*; Plenum Press: New York, **1992**.

(30) Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochem. Biophys*. **2008**, 50, 1–22.

(31) Weerasinghe, S.; Gee, M. B.; Kang, M.; Bentenitis, N.; Smith, P. E. *In Modeling Solvent Environments;* Feig, M., Ed.; Wiley-VCH: Weinheim, Germany, **2010**.

(32) Matteoli, E.; Lepori, L. *Journal of Chemical Physics* **1984**, 80, 2856–2863.

(33) Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilib*. **2010**, 290, 43–47.

(34) Chitra, R.; Smith, P. E. *Journal of Chemical Physics* **2001**, 115, 5521–5530.

(35) Hess, B.; van der Vegt, N. F. A. *Proceeding of National Academy of Science* U.S.A. **2009**, 106, 13296–13300.

(36) Klasczyk, B.; Knecht, V. *Journal of Chemical Physics* **2010**, 132, 024109.

(37) Ben-Naim, A. *Molecular Theory of Solutions*; Oxford University Press: New York, **2006**.

(38) Robinson, R. A.; Stokes, R. H. *Electrolyte Solutions*; 2nd ed.; Butterworths: London, **1959**.

(39) Rosgen, J.; Pettitt, B. M.; Perkyns, J.; Bolen, D. W. *Journal of Physical Chemistry B* **2004**, 108, 2048–2055.

(40) Sohnel, O.; Novotny, P. *Densities of Aqueous Solutions of Inorganic Substances*; Elsevier: Amsterdam, **1985**.

(41) Kirkwood, J. G.; Buff, F. P. *Journal of Chemical Physics* **1951**, 19, 774–777.

(42) Fine, R. A.; Millero, F. J. *Journal of Chemical Physics* **1973**, 59, 5529–5536.

(43) Kusalik, P. G.; Patey, G. N. *Journal of Chemical Physics* **1987**, 86, 5110–5116.

(44) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *Journal of Physical Chemistry* **1987**, 91, 6269–6271.

(45) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *Journal of Computational Chemistry* **2005**, 26, 1701–1718.

(46) Lindahl, E.; Hess, B.; van der Spoel, D. *Journal of Molecular Modeling* **2001**, 7, 306–317.

(47) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, 23, 327–341.

(48) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, 81, 3684–3690.

(49) Darden, T.; York, D.; Pedersen, L. *Journal of Chemical Physics* **1993**, 98, 10089–10092.

(50) Chandrasekhar, S. *Review of Modern Physics* **1943**, 15, 1–89.

(51) Chitra, R.; Smith, P. E. *Journal of Physical Chemistry B* **2000**, 104, 5854–5864.

(52) Allen, M. P.; Tildesley, D. J.; *Computer Simulation of Liquids*; Oxford University Press: Oxford, U. K., **1987**.

(53) Smith, P. E.; van Gunsteren, W. F. *Journal of Chemical Physics* **1994**, 100, 577–585.

(54) Walser, R.; Mark, A. E.; van Gunsteren, W. F.; Lauterbach, M.; Wipff, G. *Journal of Chemical Physics* **2000**, 112, 10450–10459.

(55) Chitra, R.; Smith, P. E. *Journal of Chemical Physics* **2001**, 114, 426–435.

(56) Kokubo, H.; Rosgen, J.; Bolen, D. W.; Pettitt, B. M. *Biophysics Journal* **2007**, 93, 3392–3407.

(57) Newman, K. E. *Chemical Society Review* **1994**, 23, 31–40.

(58) Shimizu, S.; Boon, C. L. *Journal of Chemical Physics* **2004**, 121, 9147–9155.

(59) Smith, P. E. *Fluid Phase Equilib*. **2010**, 290, 36–42.

(60) Weast, R. C. *CRC Handbook of Chemistry and Physics*; 66th ed.; CRC Press, Inc.: Boca Raton, FL, **1985**.

(61) Ansell, S.; Barnes, A. C.; Mason, P. E.; Neilson, G. W.; Ramos, S. *Biophysical Chemitry* **2006**, 124, 171–179.

(62) Marcus, Y. *Chemical Reviews* **1988**, 88, 1475–1498.

(63) Kang, M.; Smith, P. E. *Fluid Phase Equilib*. **2007**, 256, 14–19.

(64) Atkins, P. W.; De Paula, *Journal of Physical Chemistry*; 7th ed.; W.H. Freeman: New York, **2002**.

(65) Mark, P.; Nilsson, L. *Journal of Physical Chemistry A* **2001**, 105, 9954–9960.

(66) Yeh, I. C.; Hummer, G. *Journal of Physical Chemistry B* **2004**, 108, 15873–15879.

(67) Smith, P. E.; van Gunsteren, W. F. *Journal of Chemical Physics* **1994**, 100, 3169–3174.

(68) Heger, K.; Uematsu, M.; Franck, E. U. Ber. Bunsen *Physical Chemistry* **1980**, 84, 758–762.

(69) Patra, M.; Karttunen, M. *Journal of Computational Chemistry* **2004**, 25, 678–689.

(70) Davey, W. P. *Physical Review* **1923**, 21, 143–161.

(71) Wang, J. H.; Kennedy, J. W. *Journal of American Chemical Society* **1950**, 72, 2080–2083.

(72) Nelson, F.; Marcinkowsky, A. E.; Kraus, K. A. *In Research and development progress report/office of saline water;* 302nd ed.; U.S. Dept. of the Interior, Office of Saline Water: Washington, DC, **1968**.

(73) Tyrrell, H. J. V.; Harris, K. R. *Diffusion in Liquids*; Butterworths: London, **1984**.

(74) Easteal, A. J.; Woolf, L. A. *Journal of Physical Chemistry* **1986**, 90, 2441–2445.

(75) Kumamoto, E.; Kimizuka, H. Bull. *Chemical Society Janpan*. **1979**, 52, 2145–2146.

(76) Haggis, G. H.; Hasted, J. B.; Buchanan, T. J. *Journal of Chemical Physics* **1952**, 20, 1452–1465.

(77) Harris, F. E.; Okonski, C. T. *Journal of Physical Chemistry* **1957**, 61, 310–319.

(78) Buchner, R.; Hefter, G. T.; May, P. M. *Journal of Physical Chemistry A* **1999**, 103, 1–9.

(79) Beggerow, G. *In Landolt-Boernstein*; Springer-Verlag: Berlin, **1976**; Vol. 2.

(80) Matyash, I. V.; Toryanik, A. I.; Yashkichev, V. I. *Zhumal Strukturnoi Khimii* **1964**, 5, 777–778.

Appendix B - Fluctuation theory of molecular association and conformational equilibria

Yuanfang Jiao and Paul E. Smith. *Journal of Chemical Physics*, **2011**, 135, 014502

B.1 Abstract

General expressions relating the effects of pressure, temperature, and composition on solute association and conformational equilibria using the fluctuation theory of solutions are provided. The expressions are exact and can be used to interpret experimental or computer simulation data for any multicomponent mixture involving molecules of any size and character at any composition. The relationships involve particle-particle, particle-energy, and energy-energy correlations within local regions in the vicinity of each species involved in the equilibrium. In particular, it is demonstrated that the results can be used to study peptide and protein association or aggregation, protein denaturation, and protein-ligand binding. Exactly how the relevant fluctuating properties may be obtained from experimental or computer simulation data are also outlined. It is shown that the enthalpy, heat capacity, and compressibility differences associated with the equilibrium process can, in principle, be obtained from a single simulation. Fluctuation based expressions for partial molar heat capacities, thermal expansions, and isothermal compressibilities are also provided.

B.2 Introduction

Studies of protein denaturation play a central role in our efforts to understand the forces that stabilize protein structures and assemblies.[1] Proteins can be denatured by changes in temperature, pressure, and solution composition (cosolvents and pH) in closed systems[2–4] and by osmotic pressure or stress in open systems.[5] Experimentally, the thermodynamics of protein denaturation are well established and a large volume of data on protein denaturation is available. More recently, a growing amount of thermodynamic data concerning the factors that influence peptide and protein aggregation has also been determined.[6–9] Unfortunately, it is extremely difficult to relate this thermodynamic data to specific interactions with, or effects on, either the native or denatured forms. Consequently, the application of computer simulations for the study of protein denaturation has become increasingly more common. In principle, an atomic level

picture of interactions and structural changes can be elucidated from these computer simulations. However, in practice this has proven difficult for two main reasons. First, one cannot typically follow the denaturation equilibrium with current computational resources, with the possible exception of a few extreme examples,[10,11] and thereby evaluate the equilibrium constant (*K*). Second, it is not clear exactly how to extract from a simulation the relevant properties of a protein that relate to thermal or pressure denaturation—unless one has already solved the first problem.

For example, simulations of a protein folding/unfolding equilibrium to a degree where a precise equilibrium constant can be determined are extremely rare. Hence, obtaining a reliable equilibrium constant for protein denaturation over a range of pressure, temperature, or cosolvent concentrations is essentially impossible at present. The temperature denaturation or folding of small proteins or peptides can be studied more easily. In particular, peptide simulations using replica exchange techniques,[10] essentially provide the enthalpy (first derivative of *K*) and heat capacity (second derivative of *K*) changes via an analysis of the equilibrium constant as a function of temperature. However, these simulations remain computationally expensive for larger proteins in explicit solvent and it is still not clear, for instance, exactly how one should decompose or interpret the resulting enthalpy changes.[12,13] In addition, the heat capacity changes associated with thermal denaturation are also typically difficult to quantify by simulation.[14–16]

Pressure denaturation simulations are also problematic. Thermodynamics relates the effect of pressure on the equilibrium constant to a difference in volume between the native and denatured forms.[3,17] The determination of protein volumes from a simulation are either numerically challenging (direct evaluation of the volume change on addition of the protein), or require somewhat subjective definitions of the protein volume which may or may not be correct.[18–20] Furthermore, the second derivative of K with respect to pressure is usually interpreted in terms of a difference in compressibility between the two protein forms.[6] This is often estimated using a protein volume fluctuation formula which is technically only valid for the total volume of a macroscopic closed system at constant pressure and temperature.[21,22] Clearly, a more rigorous and computationally efficient approach is desirable.

The simulation of protein denaturation by the addition of cosolvents such as urea has also received attention.[23–26] In fact, the effects of cosolvents and, in particular, how these effects may be determined from simulation in a way that can be compared with experimental data are

essentially solved.[27] In our opinion, the most rigorous analysis of computer simulation data involves the Kirkwood-Buff (KB) theory of solution mixtures to relate changes in the equilibrium constant to the relative distribution of cosolvent and solvent molecules in the vicinity of both protein forms.[28–33] A similar type of approach would be useful for the interpretation of pressure and thermal denaturation simulations. This is a major goal of the present study.

The KB theory is a general theory of solution mixtures which relates molecular distributions in solution to the thermodynamic properties of that solution.[34–36] We were initially drawn to this type of approach as the resulting expressions are exact and involved quantities that can be easily obtained from computer simulations. The KB theory quantifies the molecular distributions in terms of Kirkwood-Buff integrals (KBIs), involving integrals over the corresponding intermolecular radial distribution functions, or as particle number fluctuations corresponding to local regions of the solution. Hence, it is also referred to as the fluctuation theory (FT) of solutions. The application of KB or FT to understand solution mixtures has provided valuable insight into their behavior.[37] Recently, we have extended the analysis of solution mixtures, building on the work of Buff and Brout,[38] and also Debenedetti,[39–42] by determining particle-energy and energy-energy fluctuations obtained from experimental enthalpy of mixing, thermal expansion, and heat capacity data.[43]

The KB theory has also been applied to understand chemical equilibria. Several studies have used expressions derived for thermodynamically independent infinitely dilute solutes to study the equilibrium between two infinitely dilute forms.[28,29,44,45] More rigorous work by O’Connell and co-workers has provided general expressions for reactive systems, including sequential reactions, in terms of both total and direct correlation functions after explicitly including the material balance constraint resulting in a modified grand canonical distribution function.[46,47] Ben-Naim derived expressions for the effect of a cosolvent on association equilibria in a primary solvent using an alternative approach where the chemical equilibrium conditions were imposed on the usual multicomponent KB expressions.[48] More recently, we derived general expressions which could be applied to interpret real experimental data for complex systems in a variety of ensembles using a slightly different approach, from which the original Ben-Naim result for the effect of a cosolvent could be obtained.[49] The cosolvent effects were related to particle-particle fluctuations in the vicinity of each form present in the equilibrium. Here, we wish to extend this type of approach to provide general fluctuation based expressions which can

be used to interpret the effects of temperature and pressure on association equilibria in solution, with a specific emphasis on protein denaturation. All the expressions provided are exact and can be used to interpret either experimental or simulation data concerning pressure, temperature, or cosolvent denaturation. In particular, first and second derivatives of the equilibrium constant with respect to pressure, temperature, and cosolvent concentration are developed which can (in principle) be determined from a single simulation, thereby eliminating the need for computationally intensive multiple simulations.

B.3 Theory

Here we develop relationships describing an equilibrium process in a system, which may be under a variety of different thermodynamic constraints, in terms of particle-particle, particle-energy and energy-energy fluctuations. All ensemble averages, signified in this study by angular brackets, correspond to that of the grand canonical ensemble. They can be used to describe properties of other ensembles which possess the same average thermodynamic quantities, chemical potential, pressure, etc., and then relate to fluctuations observed for local regions within these systems.

The average number density of *i* particles ($\rho_i = \langle N_i \rangle / V$) in a fixed volume V of a solution mixture can be expressed as a function of the temperature ($\beta = 1/RT$) and a set of chemical potentials ($\{\beta\mu\}$) such that,

$$d\ln\rho_i = \left(\frac{\partial \ln\rho_i}{\partial \beta}\right)_{\{\beta\mu\}} d\beta + \sum_j \left(\frac{\partial \ln\rho_i}{\partial \beta\mu_j}\right)_{\beta,\{\beta\mu\}'} d\beta\mu_j \tag{B.1}$$

where the summation is over all *j* components of the mixture, and the prime indicates that all chemical potentials except for the one of interest are held constant. Using the statistical mechanical equations associated with the grand canonical (μVT) ensemble one can show that the above derivatives are given by the following ensemble averages,[38]

$$\left(\frac{\partial \ln\rho_i}{\partial \beta\mu_j}\right)_{\beta,\{\beta\mu\}'} = \frac{<\delta N_i \delta N_j>}{<N_i>} = \delta_{ij} + N_{ij}$$

$$\left(\frac{\partial \ln\rho_i}{\partial \beta}\right)_{\{\beta\mu\}} = -\frac{<\delta N_i \delta E>}{<N_i>} = -F_{\mu,i} \tag{B.2}$$

where δ_{ij} is the Kroenecker delta function, $\delta X = X - < X >$ is the deviation in X from the ensemble average X for each member of the ensemble, and E is the total internal energy of each member of the ensemble.

The N_{ij} values represent particle-particle number fluctuations within a local volume of the solution of interest and are the focus of the KB theory of solutions. They are related to the traditional KBIs between species *i* and *j* which can be expressed in terms of the corresponding radial probability distribution (g_{ij}),[34]

$$N_{ij} = \rho_j G_{ij} = 4\pi\rho_j \int_0^\infty [g_{ij}(r) - 1] r^2 dr \tag{B.3}$$

or as particle-particle number fluctuation densities,

$$B_{ij} = \rho_i(\delta_{ij} + N_{ij}) = \frac{< \delta N_i \delta N_j >}{V} \tag{B.4}$$

The use of radial probability distributions imparts a physical picture to the N_{ij}'s. Namely, the change in the number of *j* particles resulting from the introducing an *i* particle to the reference volume, from the number of *j* particles observed in the same volume of the bulk solution.[50] The F_μ's correspond to particle number-energy correlations within the same local region of interest, and can be used to characterize solution mixtures in an analogous fashion to KB theory.[38,43]

Previously, we used a more convenient property for the analysis of experimental data on solution mixtures that provided useful relationships for particle-energy and energy-energy fluctuations in terms of experimentally accessible excess thermodynamic properties.[43] This was achieved by defining an excess energy (ε) such that,

$$\varepsilon = E - \sum_j N_j E_j^o \qquad\qquad F_i = \frac{< \delta N_i \delta \varepsilon >}{< N_i >} \tag{B.5}$$

where E_j^o is the energy per particle (molar energy) in the pure liquid *j* at the temperature and average pressure of interest. Unfortunately, when studying systems at constant pressure and temperature the above approach leads to rather cumbersome expressions. Much simpler, but totally equivalent, results can be obtained by defining an alternative fluctuating quantity such that,

$$\varepsilon_P = E - \sum_j N_j \bar{h}_j \qquad\qquad F_{P,i} = \frac{< \delta N_i \delta \varepsilon_P >}{< N_i >}$$

(B.6)

where $\bar{h}_j$ is the partial molar enthalpy of species *j* at the particular composition of interest. While it seems somewhat unusual to subtract an enthalpy term from an energy term, the following analysis is aided greatly by this substitution, especially for closed systems at constant pressure. Furthermore, while the use of partial molar enthalpies is unfortunate as they cannot be obtained from experiment, they can be easily extracted from computer simulation data, and the corresponding F_P's can be obtained from experimental data as we illustrate below. Combining Equations B.1, B.2 and B.6 one finds,

$$d\ln\rho_i = -F_{P,i} d\beta + \sum_j \left(\delta_{ij} + N_{ij}\right)\left(d\beta\mu_j - \bar{h}_j d\beta\right)$$

(B.7)

for any species *i* in a mixture of *j* components at any composition. If T (β) is constant the above equations reduce to a series of differentials corresponding to the traditional KB theory of solution mixtures.[51,52] Using the above substitution one finds that the isothermal compressibility (κ_T), isobaric thermal expansion coefficient (α_P), and constant pressure heat capacity (C_P) of any multicomponent solution mixture are then given by,

$$RT\kappa_T = -RT\left(\frac{\partial \ln V}{\partial P}\right)_{T,\{N\}} = \sum_j (\delta_{ij} + N_{ij})\bar{V}_j$$

$$RT^2\alpha_P = RT^2\left(\frac{\partial \ln V}{\partial T}\right)_{P,\{N\}} = -F_{P,i} = -\sum_i x_i F_{P,i} = -\sum_i \phi_i F_{P,i}$$

$$RT^2 C_P = RT^2\left(\frac{\partial <\mathrm{H}>}{\partial T}\right)_{P,\{N\}} = < \delta\varepsilon_P \delta\varepsilon_P >$$

(B.8)

where $\bar{V}_i$ and $\phi_i = \rho_i \bar{V}_i$ are the partial molar volume and volume fraction of *i*, respectively. The above expressions are much simpler than the equivalent expression provided previously for the thermal expansion and heat capacity.[38,43,53]

In our later analysis we shall also use the pseudo chemical potential (μ_i^*) concept, and its associated enthalpy (h_i^*) and volume (V_i^*), to indicate how one can extract relevant quantities from available experimental data. The pseudo chemical potential approach centers on the

statistical mechanical definition of chemical potential and thereby eliminates the need for standard states.[36] The pseudo chemical potential is similar to the excess chemical potential used in computer simulations with the only difference relating the internal partition function of the species. Using Equations B.1 and B.2 and rearranging one finds,

$$d\beta\mu_i^* \equiv d\beta\mu_i - d\ln\Lambda_i^3\rho_i = -3d\ln\Lambda_i + F_{\mu,i}d\beta - \sum_j N_{ij}d\beta\mu_j \tag{B.9}$$

where Λ_i is the thermal de Broglie wavelength and is proportional to $T^{-1/2}$. From the above equation one obtains,

$$V_i^* = \bar{V}_i - RT\kappa_T = -\sum_j N_{ij}\bar{V}_j$$

$$h_i^* = \bar{h}_i - RT^2\alpha_P - \tfrac{3}{2}RT = -\tfrac{3}{2}RT + F_{\mu,i} - \sum_j N_{ij}\bar{h}_j$$

$$d\mu_i^* = d\mu_i - RTd\ln\rho_i = -\sum_j N_{ij}d\mu_j \qquad \text{T constant}$$

(B.10)

where the last equation can be used for changes in composition of the system. The volume term (V_i^*) can be determined experimentally, while only changes in the enthalpy and chemical potentials for various processes can be evaluated experimentally (see later discussion). In principle, all three properties can be obtained directly from computer simulations.

Equation B.7 represents a series of source equations which can be used to obtain expressions for various properties of solution mixtures in terms of number-number, number-energy, and (later) energy-energy correlations characterizing local microscopic regions within the solution. Our primary focus here is that of chemical equilibria involving an associating solute, or a solute which can undergo a change in conformation. We examine a system with a solute (2) in a primary solvent (1) which may contain any number of additional cosolvents (3,4, …). The solute is in equilibrium between two forms. One form being a monomer (M) and the other an aggregate (A) containing *n* monomers. This equilibrium is described by an equilibrium constant (K) such that,

$$nM \rightarrow A \qquad K = \frac{\rho_A}{\rho_M^n} \qquad d\ln K = d\ln\rho_A - nd\ln\rho_M \tag{B.11}$$

We note that the equilibrium constant involves the actual number densities (molar concentrations) present at the equilibrium composition of interest, and not the activities approximated by concentrations – as is often the case in biological problems. The relationships between the number of solutes, number of monomers, and number of aggregates are given by,

$$N_M + nN_A = N_2 \qquad f_M = \frac{N_M}{N_2} \qquad f_A = \frac{nN_A}{N_2} \tag{B.12}$$

Furthermore, the equilibrium conditions dictate that the flowing relationships,

$$\mu_A = n\mu_M = n\mu_2 \qquad d\mu_A = nd\mu_M = nd\mu_2 \tag{B.13}$$

must be obeyed. Using Equation B.7, the above relationships, and the approach outlined previously,[49] it is relatively easy to show that,

$$d\ln K = -\left(F_{P,A} - nF_{P,M}\right)d\beta + \sum_i (N_{Ai} - nN_{Mi})(d\beta\mu_i - \bar{h}_i d\beta) \tag{B.14}$$

where the summation is over all thermodynamically independent i components including component 2. The exact meaning of the N_{A2} and N_{M2} values has been discussed previously.[49] This equation, in combination with the Gibbs-Duhem (GD) expression,

$$SdT - VdP + \sum_i N_i d\mu_i = 0 \tag{B.15}$$

enables one to develop a complete picture of how pressure, temperature, and solution composition affect the above equilibrium in terms of particle and energy fluctuations within local regions of the solution under the conditions of interest.

B.4 Results

In the following sections we provide expressions (first and second derivatives) describing the effect of pressure, temperature and composition on a general equilibrium process in solution. The equilibrium can involve molecules of any size and character. In addition, we provide expressions for the simplest case – a two state conformational equilibrium in a single solvent at infinite dilution of the solute – such as often used to understand protein folding or denaturation.

B.4.1 The Effect of Hydrostatic Pressure on Chemical Equilibria

Taking derivatives of Equation 14 with respect to pressure while keeping temperature and composition constant immediately provides,

$$\left(\frac{\partial \ln K}{\partial P}\right)_{\beta,\{N\}} = \beta \sum_i (N_{Ai} - nN_{Mi})\overline{V}_i \tag{B.16}$$

The partial molar volumes can also be expressed in terms of KBIs if desired. Hence, in the absence of specific affinities between the two forms and any *i* species, then if A is "smaller" than *n* monomers there will be an excess of each *i* molecule in the vicinity of A compared to the vicinity of *n* monomers. Therefore, each term on the right hand side will be positive and an increase in P will increase K and thus favor the A form.

In many cases one is also interested in the "compressibility" of the process as manifested in the second derivative of the equilibrium constant with respect to pressure. To develop expressions for this derivative we first note that,

$$\left(\frac{\partial \ln K}{\partial P}\right)_{\beta,\{N\}} = \beta \sum_i \overline{V}_i \left(\frac{\partial \ln K}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'} \tag{B.17}$$

Consequently, taking pressure derivatives of the above expression, then interchanging the order of differentiation, one finds the second derivative can be written as,

$$\left(\frac{\partial^2 \ln K}{\partial P^2}\right)_{\beta,\{N\}} = -\beta \sum_i (N_{Ai} - nN_{Mi})\bar{V}_i \bar{\kappa}_{T,i} + \beta^2 \sum_{i,j} \bar{V}_i \bar{V}_j \left(\frac{\partial (N_{Aj} - nN_{Mj})}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'} \tag{B.18}$$

where $\bar{\kappa}_{T,i} = -(\partial \ln \bar{V}_i / \partial P)_{T,\{N\}}$ is the partial molar isothermal compressibility of species *i*, and $(\partial \bar{V}_j / \partial \beta \mu_i)_{\beta,\{\beta\mu\}'} = 0$. Fluctuation based expressions for the partial molar compressibilities are provided in Appendix 2. The required derivative can be obtained from the equations of the grand canonical ensemble in the same manner as before to provide,

$$\left(\frac{\partial N_{Aj}}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'} = \frac{< \delta N_A \delta N_i \delta N_j >}{< N_A >} - N_{Ai} N_{Aj} = N_{Aij} - N_{Ai} N_{Aj} \tag{B.19}$$

Hence, the above derivative provides information on triplet particle number fluctuations in the region of interest. Similar expressions are obtained for the corresponding derivatives of N_{Mj}. The

final expression for the second derivative of the equilibrium with respect to pressure for a solution containing any number of components is therefore,

$$\left(\frac{\partial^2 \ln K}{\partial P^2}\right)_{\beta,\{N\}} = -\beta \sum_i (N_{Ai} - nN_{Mi})\bar{V}_i \bar{\kappa}_{T,i} + \beta^2 \sum_{i,j} \bar{V}_i \bar{V}_j \left[N_{Aij} - N_{Ai}N_{Aj} - n\left(N_{Mij} - N_{Mi}N_{Mj}\right)\right]$$

(B.20)

In principle, all the terms present in Equations B.16 and B.20 can be determined reasonably easily from a single computer simulation.

In order to extract the specific terms associated with each form (A or M) one requires knowledge of the solute volume and how it varies with pressure. From Equation B.10 we have,

$$V_2^* = -\sum_j N_{2j}\bar{V}_j = -\sum_j (f_A N_{Aj} + f_M N_{Mj})\bar{V}_j$$

(B.21)

which provides the individual sums over all j terms. To extract each N_{Aj} term would require knowledge of all the partial molar volumes in the mixture. The pressure derivative of the solute volume can then be expressed as,

$$\left(\frac{\partial V_i^*}{\partial P}\right)_{\beta,\{N\}} = -\sum_j \left[\left(\frac{\partial N_{ij}}{\partial \mathrm{P}}\right)_{\beta,\{N\}} \overline{V}_j - N_{ij}\bar{V}_j \bar{\kappa}_{T,j}\right]$$

(B.22)

where the required derivative is given by,

$$\left(\frac{\partial N_{ij}}{\partial P}\right)_{\beta,\{N\}} = -N_{ij}\kappa_T + \beta \sum_k \frac{< \delta N_i \delta N_j \delta N_k >}{< N_i >} \bar{V}_k$$

(B.23)

and was evaluated by treating $< \delta N_i \, \delta N_j >$ as a function of $\{\beta\mu\}$, β, and V, in a similar manner to Equation 1 (see Appendix 1). The final expression for the change in solute volume with pressure is then,

$$\left(\frac{\partial V_2^*}{\partial P}\right)_{\beta,\{N\}} = RT(\kappa_T)^2 + \sum_i N_{2i}\bar{V}_i \bar{\kappa}_{T,i} - \beta \sum_{i,j} \bar{V}_i \bar{V}_j \left[f_A N_{Aij} + f_M N_{Mij}\right]$$

(B.24)

We note that Equation B.16, B.20, and B.21 could be simplified by using the notation $V_A^* = -\sum_j N_{Aj}\bar{V}_j$, as suggested by Equation 10, although this requires some care (see later discussion).

A situation of common interest is that of an infinitely dilute solute in a single solvent. For instance, experimental data concerning the pressure denaturation (N→ D) of proteins is often interpreted in terms of a Taylor expansion of the standard free energy change for unfolding,[6]

$$-\beta\Delta\Delta G^{o,\infty} = \ln\left(\frac{K}{K_0}\right)^{\infty} \approx \left(\frac{\partial \ln K}{\partial P}\right)^{\infty}_{T,m_2} \Delta P + \frac{1}{2!}\left(\frac{\partial^2 \ln K}{\partial P^2}\right)^{\infty}_{T,m_2} (\Delta P)^2 + O[(\Delta P)^3]$$

(B.25)

where the subscript 0 refers to the reference pressure, $\Delta P = P - P_0$, and all derivatives are evaluated at P_0 (usually 1 bar). In this case the first derivative is expressed in terms of fluctuating quantities as,

$$\left(\frac{\partial \ln K}{\partial P}\right)^{\infty}_{T,m_2} = \beta(G^{\infty}_{D1} - G^{\infty}_{N1}) = \beta V_1^o(< N_1 >_D - < N_1 >_N) = -\beta\Delta V^{o,\infty}$$

(B.26)

where $m_2 = \rho_2/\rho_1$, is a dimensionless molality, and the last equality arises from standard thermodynamics under conditions that activity and concentration are equivalent. The above expression has been presented before,[28] although it is seldom used for the analysis of simulation data on pressure denaturation.[54] The subscript D (or N) indicates an ensemble average within the same fixed volume of solution surrounding a single D (or N) molecule. Hence, the KBIs essentially quantify the volume of each solute form in terms of the number of solvent molecules that can be accommodated in the same fixed volume of solution.

The second derivative for pressure denaturation is given by,

$$\left(\frac{\partial^2 \ln K}{\partial P^2}\right)^{\infty}_{T,m_2} = -\beta(G^{\infty}_{D1} - G^{\infty}_{N1})\kappa_T^o + (\beta V_1^o)^2\left[< \delta N_1 \delta N_1 >_D - < \delta N_1 \delta N_1 >_N - [(N^{\infty}_{D1})^2 - (N^{\infty}_{N1})^2]\right]$$

(B.27)

It is comforting to note the similarity of several terms in the above expression to the compressibility equation for a pure solvent, involving fluctuations in the number of solvent particles.[36] The above expressions indicate that the volume and "compressibility" associated with the equilibrium, and thereby the proteins themselves, are actually properties solely related to the water distribution in the vicinity of each form. In the present fluctuation based approach the protein volume does not enter directly into the expressions, and hence one does not have to be concerned as to exactly how this is defined or represented when analyzing computer simulation

data. However, clearly the number of waters and their fluctuations will mimic the protein volume and fluctuations in the protein volume (see later discussion).

In order to extract specific fluctuations associated with each form one requires information concerning the protein solute volume and how the volume changes with pressure. Namely,

$$V_2^{*,\infty} = -G_{21}^{\infty} = -f_D G_{D1}^{\infty} - f_N G_{N1}^{\infty} \tag{B.28}$$

and,

$$\left(\frac{\partial V_2^*}{\partial P}\right)_{T,m_2}^{\infty} = RT(\kappa_T^0)^2 + G_{21}^{\infty}\kappa_T^0 - \beta(V_1^0)^2[f_D < \delta N_1 \delta N_1 >_D + f_N < \delta N_1 \delta N_1 >_N - < \delta N_1 \delta N_1 >_0] \tag{B.29}$$

The zero subscript indicating an ensemble average obtained for the same volume of pure solvent. Hence, if one knows K and the compressibility of the pure solvent, together with $V_2^{*,\infty}$ for a series of pressures of interest, then the individual KBIs and fluctuations can be extracted from experimental data.

B.4.2 The Effect of Temperature on Chemical Equilibria at Constant Pressure

One of the most common ways to affect a chemical equilibrium involves changes in temperature at constant pressure. To our knowledge there are no fluctuation based expressions currently available for describing the associated enthalpy and heat capacity changes for chemical equilibria. Taking derivatives of Equation 14 with respect to β keeping pressure and composition constant one immediately obtains,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{P,\{N\}} = -\left(F_{P,A} - nF_{P,M}\right) \tag{B.30}$$

The above expression is valid for any number of components at any concentration. In order to develop second derivatives of the equilibrium constant with respect to β we first note that,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{P,\{N\}} = \left(\frac{\partial \ln K}{\partial \beta}\right)_{\{\beta\mu\}} + \sum_i \bar{h}_i \left(\frac{\partial \ln K}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'}$$

(B.31)

The second derivative is then obtained from the derivative of Equation B.31 after a change in the order of differentiation,

$$\left(\frac{\partial^2 \ln K}{\partial \beta^2}\right)_{P,\{N\}} = -RT^2 \sum_i (N_{Ai} - nN_{Mi})\, \bar{c}_{P,i} - \left(\frac{\partial\left(F_{P,A} - nF_{P,M}\right)}{\partial \beta}\right)_{\{\beta\mu\}} - \sum_i \bar{h}_i \left(\frac{\partial\left(F_{P,A} - nF_{P,M}\right)}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'}$$

(B.32)

where $\bar{c}_{p,i}$ is the partial molar constant pressure heat capacity of species *i* (see Appendix 2 for the relevant expressions). All the derivatives in the above expression can be evaluated in terms of local fluctuations using the equations of the grand canonical ensemble. Noting that $(\partial \bar{h}_j / \partial \beta \mu_i)_{\beta,\{\beta\mu\}'} = (\partial \bar{h}_j / \partial \beta)_{\{\beta\mu\}} = 0$, the results are given by the following derivatives,

$$\left(\frac{\partial F_{P,A}}{\partial \beta}\right)_{\{\beta\mu\}} = -\frac{< \delta N_A \delta \varepsilon_P \delta E >}{< N_A >} + F_{P,A} F_{\mu,A}$$

(B.33)

and,

$$\left(\frac{\partial F_{P,A}}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'} = \frac{< \delta N_A \delta N_i \delta \varepsilon_P >}{< N_A >} - F_{P,A} N_{Ai}$$

(B.34)

which together provide,

$$\left(\frac{\partial F_{P,A}}{\partial \beta}\right)_{\{\beta\mu\}} + \sum_i \bar{h}_i \left(\frac{\partial F_{P,A}}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'} = -\frac{< \delta N_A \delta \varepsilon_P \delta \varepsilon_P >}{< N_A >} + F_{P,A}^2$$

(B.35)

Similar expressions are found for the monomer form. Hence, the final result for a solution containing any number of components is given by,

$$\left(\frac{\partial^2 \ln K}{\partial \beta^2}\right)_{P,\{N\}} = -RT^2 \sum_i (N_{Ai} - nN_{Mi}) \bar{c}_{P,i} + \frac{< \delta N_A \delta \varepsilon_P \delta \varepsilon_P >}{< N_A >} - (F_{P,A})^2 - n\left[\frac{< \delta N_M \delta \varepsilon_P \delta \varepsilon_P >}{< N_M >} - (F_{P,M})^2\right]$$

(B.36)

Again, it is satisfying that the above expression involves local energy fluctuations in the vicinity of both forms of the solute, which are typically characteristic of heat capacities.[36]

In order to extract the specific terms for each form one requires knowledge of the thermal expansion and how the pseudo enthalpy varies with temperature. From Equation B.8 one has,

$$RT^2\alpha_P = -F_{P,2} = -f_A F_{P,A} - f_M F_{P,M} \tag{B.37}$$

and from Equation B.10,

$$\left(\frac{\partial h_i^*}{\partial \beta}\right)_{P,\{N\}} = \tfrac{3}{2}(RT)^2 + \left(\frac{\partial F_{\mu,i}}{\partial \beta}\right)_{P,\{N\}} - \sum_j \left[\bar{h}_j \left(\frac{\partial N_{ij}}{\partial \beta}\right)_{P,\{N\}} - RT^2 N_{ij}\bar{c}_{P,j}\right] \tag{B.38}$$

The two constant pressure derivatives can be evaluated using the same approach as used for Equation B.23 (see Appendix 1) to provide,

$$\begin{aligned}\left(\frac{\partial h_2^*}{\partial \beta}\right)_{P,\{N\}} &= -RT^2 c_{p,2}^* \\ &= \tfrac{3}{2}(RT)^2 + (RT^2\alpha_P)^2 - f_A \frac{< \delta N_A \delta\varepsilon_P \delta\varepsilon_P >}{< N_A >} - f_M \frac{< \delta N_M \delta\varepsilon_P \delta\varepsilon_P >}{< N_M >} \\ &\quad + RT^2 \sum_j N_{2j}\bar{c}_{P,j}\end{aligned} \tag{B.39}$$

Analogous expressions can be obtained for the other solution components if required.

Focusing again on an infinitely dilute protein solute in a single solvent one finds a Taylor expansion of the equilibrium constant for a simple two state denaturation around a reference temperature provides the usual relationship,

$$\ln\left(\frac{K}{K_0}\right)^{\infty} \approx \left(\frac{\partial \ln K}{\partial T}\right)_{P,m_2}^{\infty} \Delta T + \frac{1}{2!}\left(\frac{\partial^2 \ln K}{\partial T^2}\right)_{P,m_2}^{\infty} (\Delta T)^2 + O[(\Delta T)^3] \tag{B.40}$$

where the first derivative is given by,

$$\left(\frac{\partial \ln K}{\partial T}\right)_{P,m_2}^{\infty} = (RT^2)^{-1}(F_{P,D}^{\infty} - F_{P,N}^{\infty}) = (RT^2)^{-1}\Delta H^{o,\infty} \tag{B.41}$$

and provides an expression for the standard enthalpy change in terms of fluctuations in the local solution properties. This could be particularly useful for the analysis of computer simulation data. The value of $F_{P,D}$ at infinite dilution is given by,

$$F_{P,D}^{\infty} =< \varepsilon_P >_D -< \varepsilon_P >_o =< E_{DD} + E_{D1} + E_{11} - N_1 H_1^o >_D -< E_{11} - N_1 H_1^o >_o$$

(B.42)

which contains terms one would expect, namely the intra and intermolecular energies E_{ij}, together with a less obvious additional term $N_1 H_1^o$, which is non negligible. The last term in angular brackets will cancel when n = 1. The above result can be written in a more noticeable form using an alternative energy fluctuation term,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{P,m_2}^{\infty} = -(F_D^{\infty} - F_N^{\infty}) + P(G_{D1}^{\infty} - G_{N1}^{\infty}) = -\Delta E^{o,\infty} - P\Delta V^{o,\infty}$$

(B.43)

where we have used the definition of ε presented in Equation 5. The volume term is the same as that derived for pressure denaturation (Equation 26), while the energy term provides a fluctuation expression for the energy contribution to the enthalpy change. The second derivative of the equilibrium constant is given by,

$$\begin{aligned}\left(\frac{\partial^2 \ln K}{\partial T^2}\right)_{P,m_2}^{\infty} &= (RT^2)^{-1}\Delta C_p^o \\ &= -(RT^2)^{-1} C_{p,m}^o (N_{D1}^{\infty} - N_{N1}^{\infty}) \\ &+ (RT^2)^{-2}\left[< \delta\varepsilon_P \delta\varepsilon_P >_D -< \delta\varepsilon_P \delta\varepsilon_P >_N - [(F_{P,D}^{\infty})^2 - \left(F_{P,N}^{\infty}\right)^2]\right]\end{aligned}$$

(B.44)

and provides a route to heat capacity changes associated with the chemical equilibrium in terms of local fluctuations. In the majority of cases the volume changes associated with the equilibrium will be negligible in solution and it is often safe to ignore the PV_i terms (≈2 J/mol for water) and use $\varepsilon_P = \varepsilon$ in Equations [41-44].

Given an expression for the pseudo enthalpy of an infinitely dilute solute in a single solvent provided by Equation B.10 we have,

$$h_2^{*,\infty} = -\frac{3}{2}RT + F_{\mu,2}^{\infty} - N_{21}^{\infty} H_1^o = -\frac{3}{2}RT + F_{P,2}^{\infty} = -\frac{3}{2}RT + f_A F_{P,D}^{\infty} + f_M F_{P,N}^{\infty}$$

(B.45)

and from Equation B.39,

$$\left(\frac{\partial h_2^*}{\partial \beta}\right)_{P,m_2}^{\infty} = -RT^2 c_{p,2}^{*,\infty}$$
$$= \tfrac{3}{2}(RT)^2 + (RT^2\alpha_P^0)^2 + RT^2 N_{21}^{\infty} C_{P,m}^0$$
$$- [f_D < \delta\varepsilon_P \delta\varepsilon_P >_D + f_N < \delta\varepsilon_P \delta\varepsilon_P >_N - < \delta\varepsilon_P \delta\varepsilon_P >_o]$$

(B.46)

Hence, if one knows K, the thermal expansion and heat capacity of the pure solvent, together with $h_2^{*,\infty}$ for a relevant process over a series of temperatures of interest, then the individual F_P's and fluctuations in ε_P can be extracted from experimental data.

Before leaving this section we note that occasionally the pressure-temperature cross derivative may be useful and can be obtained from Equation 16 to give,

$$\left(\frac{\partial}{\partial \beta}\left(\frac{\partial \ln K}{P}\right)_{\beta,m_2}\right)_{P,m_2}^{\infty}$$
$$= (G_{D1}^{\infty} - G_{N1}^{\infty})(1 - T\alpha_P^0)$$
$$- \beta V_1^0 \left[< \delta N_1 \delta\varepsilon_P >_D - < \delta N_1 \delta\varepsilon_P >_N - [F_{P,D}^{\infty} N_{D1}^{\infty} - F_{P,N}^{\infty} N_{N1}^{\infty}]\right]$$

(B.47)

for protein denaturation at infinite dilution.

B.4.3 The Effect of Temperature on Chemical Equilibria at Constant Volume

Alternatively, the change in temperature could be performed for an equilibrium process under conditions of constant volume. While this is not a common situation, it is included here for completeness. In this case the general expression becomes,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{V,\{N\}} = -\left(F_{P,A} - nF_{P,M}\right) - \frac{T\alpha_P}{\kappa_T}\sum_i (N_{Ai} - nN_{Mi})\bar{V}_i$$

(B.48)

where we have used the following thermodynamic relationship,

$$\left(\frac{\partial \beta\mu_i}{\partial \beta}\right)_{V,\{N\}} = \bar{h}_i - \bar{V}_i \frac{T\alpha_P}{\kappa_T}$$

(B.49)

All the terms in the above equation represent average properties of the solution mixture and not fluctuating quantities. The change in equilibrium constant can be expressed in a far simpler form if one defines an alternative fluctuating property. We choose,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{V,\{N\}} = -\left(F_{V,A} - nF_{V,M}\right)$$

(B.50)

where,

$$F_{V,i} = \frac{< \delta N_i \delta \varepsilon_V >}{< N_i >} \qquad \varepsilon_V = E - \sum_j N_j \left(\bar{h}_j - \bar{V}_j \frac{T\alpha_P}{\kappa_T}\right)$$

(B.51)

If required, the thermal expansion and compressibility can be expressed in terms of fluctuating quantities. It should also be noted that using this definition one can show that,

$$RT^2 C_V = RT^2 \left(\frac{\partial <\mathrm{E}>}{\partial T}\right)_{V,\{N\}} = < \delta \varepsilon_V \delta \varepsilon_V >$$

(B.52)

which is much simpler than previous expressions for C_v.[38,53]

Second derivatives of the equilibrium constant will eventually lead to an expression for the constant volume heat capacity change associated with the equilibrium. First, we note that,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{V,\{N\}} = \left(\frac{\partial \ln K}{\partial \beta}\right)_{\{\beta\mu\}} + \sum_i \left(\bar{h}_i - \bar{V}_i \frac{T\alpha_P}{\kappa_T}\right)\left(\frac{\partial \ln K}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'}$$

(B.53)

and hence,

$$\left(\frac{\partial^2 \ln K}{\partial \beta^2}\right)_{V,\{N\}} = -RT^2 \sum_i (N_{Ai} - nN_{Mi})\, \bar{c}_{V,i} - \left(\frac{\partial \left(F_{V,A} - nF_{V,M}\right)}{\partial \beta}\right)_{\{\beta\mu\}} - \sum_i \left(\bar{h}_i - \bar{V}_i \frac{T\alpha_P}{\kappa_T}\right)\left(\frac{\partial \left(F_{V,A} - nF_{V,M}\right)}{\partial \beta \mu_i}\right)_{\beta,\{\beta\mu\}'}$$

(B.54)

using the same approach as before. Fluctuation based expressions for $\bar{c}_{V,i}$ can be found in Appendix 2. The derivatives in the above expression are analogous to Equations 33-35 and lead to the final result,

$$\left(\frac{\partial^2 \ln K}{\partial \beta^2}\right)_{V,\{N\}} = -RT^2 \sum_i (N_{Ai} - nN_{Mi}) \bar{c}_{V,i} + \frac{< \delta N_A \delta \varepsilon_V \delta \varepsilon_V >}{< N_A >} - (F_{V,A})^2 - n\left[\frac{< \delta N_M \delta \varepsilon_V \delta \varepsilon_V >}{< N_M >} - (F_{V,M})^2\right]$$

(B.55)

which takes the same form as the constant P expression, although the fluctuating quantities are clearly different.

B.4.4 The Effect of Cosolvents on Chemical Equilibria

The effect of cosolvents on chemical equilibria has been the subject of many KB related studies.[27] Recently, we provided a general multicomponent expression in terms of chemical potential derivatives.[49] A similar expression, but using molarities instead of molalities, can be obtained from Equation 14 after taking derivatives with respect to the (molar) concentration of species *j* keeping pressure, temperature, and the number of all other species constant,

$$\left(\frac{\partial \ln K}{\partial \ln \rho_j}\right)_{P,T,\{m\}'} = \sum_i (N_{Ai} - nN_{Mi})\mu_{ij} = \sum_{i\neq 1} (\Gamma_{Ai} - n\Gamma_{Mi})\mu_{ij}$$

(B.56)

where $\mu_{ij} = (\partial \mu_i / \partial \ln \rho_j)_{P,T,\{m\}'}$ and the last summation excludes the primary solvent after elimination of $d\mu_1$ using the GD expression at constant T and P. The Γs are defined by,

$$\Gamma_{Aj} = \rho_j \left(G_{Aj} - G_{A1}\right) = N_{Aj} - m_j N_{A1}$$

(B.57)

and can be considered as preferential binding parameters which quantify the excess binding of species *j* relative to that of the primary solvent in the vicinity of each solute form. The chemical potential derivatives can also be expressed in terms of KBIs if desired.[34,51] For an infinitely dilute solute in a primary solvent the addition of a single cosolvent (3) results in a change to the equilibrium provided by,[29,45]

$$\left(\frac{\partial \ln K}{\partial \ln \rho_3}\right)^{\infty}_{P,T,m_2} = (\Gamma^{\infty}_{A3} - \mathrm{n}\Gamma^{\infty}_{M3})\mu_{33}$$

(B.58)

with,

$$\mu_{33} = \beta \left(\frac{\partial \mu_3}{\partial \ln \rho_3}\right)^{\infty}_{P,T,m_2} = \frac{1}{1 + N_{33} - N_{13}}$$

(B.59)

Note that the above derivative is different (molarity versus molality) than used in the traditional notation. Derivatives using other concentration scales can be found using the relationships provided by standard thermodynamics and KB theory,

$$\left(\frac{\partial \ln \rho_3}{\partial \ln m_3}\right)^{\infty}_{P,T,m_2} = \phi_1 = \frac{1 + N_{33} - N_{13}}{1 + N_{33}^{+}} \qquad \left(\frac{\partial \ln x_3}{\partial \ln m_3}\right)^{\infty}_{P,T,m_2} = x_1$$

(B.60)

with $N_{ij}^{+} = N_{ij} + m_j(1 + N_{11} - N_{i1} - N_{j1})$. The above result (Equation B.58) was also obtained by Ben-Naim using a different approach and alternative definitions of the equilibrium constant and cosolvent concentration.[48]

In order to isolate the binding to either form one can take one of two approaches. First, rearranging Equation B.10, then eliminating $d\mu_1$ using the GD equation at constant T and P, provides an expression valid only for systems at constant temperature and pressure,

$$-d\mu_i^{*} = -RTd\ln y_i = \sum_{j \neq 1} \Gamma_{ij} d\mu_j$$

(B.61)

where y_i is the molar activity coefficient and μ_i^* is the pseudo chemical potential of species *i* (≠ 1). Derivatives of this expression then provide the following,

$$-\left(\frac{\partial \mu_2^{*}}{\partial \mu_3}\right)^{\infty}_{P,T,m_2} = -\left(\frac{\partial \ln y_2}{\partial \ln a_3}\right)^{\infty}_{P,T,m_2} = \Gamma_{23}^{\infty} = f_A \Gamma_{A3}^{\infty} + f_M \Gamma_{M3}^{\infty}$$

(B.62)

Therefore, experimental data concerning the behavior of the solute activity coefficient at low concentrations can be used to extract values for Γ_{23}, etc. Alternatively, equilibrium dialysis studies that measure density changes in osmotic systems on the introduction of a non-diffusible solute (such as a protein) provide,[55]

$$\left(\frac{\partial m_3}{\partial m_2}\right)^{\infty}_{T,\mu_1,\mu_3} = \Gamma_{23}^{\infty} = f_A \Gamma_{A3}^{\infty} + f_M \Gamma_{M3}^{\infty}$$

(B.63)

and also enable the isolation of the various Γ values. Finally, before leaving this section we note that typical cosolvent denaturation studies monitor the change in equilibrium constant as a function of cosolvent molarity,

$$\left(\frac{\partial \ln K}{\partial \rho_3}\right)^{\infty}_{P,T,m_2} = \frac{1}{\rho_3}\frac{\Gamma^{\infty}_{D3} - \Gamma^{\infty}_{N3}}{1 + N_{33} - N_{13}} = m \tag{B.64}$$

from which a fluctuation based expression for the protein *m*-value is obtained.[56]

It is possible to determine second derivatives of the equilibrium constant with respect to cosolvent concentration. Indeed, nonlinear cosolvent effects are observed.[57] The general expression is,

$$\rho_j^2\left(\frac{\partial^2 \ln K}{\partial \rho_j^2}\right)_{P,T,\{m\}'} = \sum_i (N_{Ai} - nN_{Mi})\mu_{ijj} + \sum_{i,k}\mu_{ij}\mu_{kj}[N_{Aik} - N_{Ai}N_{Ak} - n(N_{Mik} - N_{Mi}N_{Mk})] \tag{B.65}$$

where $\mu_{ijj} = (\partial \mu_{ij}/\partial \ln \rho_j)_{P,T,\{m\}'}$. Some of the above terms can be eliminated using the GD equation and Equation 10. The most useful relationship is obtained for the denaturation of an infinitely dilute protein solute in a mixture of solvent (1) and a single cosolvent (3). Here, one finds,

$$\rho_3^2\left(\frac{\partial^2 \ln K}{\partial \rho_3^2}\right)^{\infty}_{P,T,m_2} = \sum_{i=1,3}(N_{Ai} - nN_{Mi})\mu_{i33} + \mu_{33}^2\left[< \delta\Gamma^{\infty}_{D3}\delta\Gamma^{\infty}_{D3} >_D - < \delta\Gamma^{\infty}_{N3}\delta\Gamma^{\infty}_{N3} >_N - [(\Gamma^{\infty}_{D3})^2 - (\Gamma^{\infty}_{N3})^2]\right] \tag{B.66}$$

where $\Gamma^{\infty}_{D3} =< N_3 >_D - m_3 < N_1 >_D$ and $\delta\Gamma^{\infty}_{D3} = \delta N_3 - m_3\delta N_1$, which corresponds to a fluctuation in the binding parameter, i.e. cosolvent and water distributions, in the vicinity of a single denatured form. To our knowledge, general fluctuation expressions for the μ_{ijj} derivatives are not available. However, based on calculations described in the Discussion section we suspect the contribution from the first term on the right hand side will be negligible.

B.4.5 Osmotic Systems

Many equilibria of biological importance occur under osmotic (or cellular) conditions. In addition, the study of osmotic pressures in protein solutions is interesting in that it provides information concerning protein-protein interactions.[58] An expression for the change in osmotic pressure (Π) on the addition of a biomolecule can be obtained from Equations B.7 and B.15,

$$\left(\frac{\partial \Pi}{\partial \rho_2}\right)_{\beta,\{\beta\mu\}'} = \frac{RT}{1+N_{22}}$$

(B.67)

and was noted in the original KB study.[34] This expression is valid for systems containing any number of additional cosolvents as long as their chemical potentials are held constant. Higher derivatives of the osmotic pressure provide information on higher order correlations between solute molecules. For instance, the second derivative of the osmotic pressure is given by,

$$\left(\frac{\partial^2 \Pi}{\partial {\rho_2}^2}\right)_{\beta,\{\beta\mu\}'} = \frac{RT}{\rho_2(1+N_{22})}\left[1 - \frac{<\delta N_2 \delta N_2 \delta N_2>}{<N_2>(1+N_{22})^2}\right]$$

(B.68)

and includes information concerning triplet solute correlations. Kirkwood and Buff showed that these higher order terms can form part of a series expansion which reduces to the McMillan-Mayer theory of solutions when the derivatives are obtained at infinite dilution of the solute.[22] However, the above expressions are valid for any solution composition. For chemical equilibria in an osmotic system one finds from Equation 14 that,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)^{\infty}_{\rho_2,\{\beta\mu\}'} = -\left(F^{\infty}_{P,A} - nF^{\infty}_{P,M}\right) - \left(N^{\infty}_{A1} - nN^{\infty}_{M1}\right)H^{o}_{1} = -\left(F^{\infty}_{\mu,A} - nF^{\infty}_{\mu,M}\right)$$

(B.69)

for an infinitely dilute solute in an osmotic solution. This result is to be expected as the conditions are essentially those of the grand canonical ensemble. Second derivatives can be obtained and provide the same expression as found in Equation B.44, but where ε_P is replaced by E and F_P is replaced by F_μ, and there is no term involving the heat capacity.

B.5 Discussion

The expressions provided here can be used to interpret experimental data on chemical equilibria. In particular, the individual terms can be extracted under favorable conditions where the relevant experimental data is available. Alternatively, they can be used to analyze computer simulation data. In the latter case fluctuations of the form $< \delta X\ \delta Y\ \delta Z >$ need to be evaluated from the trajectory. In principle, these can be expressed in terms of correlation functions in a similar way to the usual KBIs.[34] However, in practice this is actually more cumbersome than evaluating the fluctuations within the local regions directly. The expressions are all exact. We are

currently using computer simulations and applying Equations B.26, B.41 and B.58 to study the effects of temperature, pressure and composition on a simple conformational equilibrium in pure water.[59] The numerical results support the validity of the expressions presented here. We also note that, while primarily developed for the description of solution mixtures, the expressions are also valid for gaseous systems.

One of the more interesting results obtained in this study is the expression given in Equation B.41 which provides a route to the standard enthalpy change for association or denaturation at infinite dilution of the solute. It is also clear from Equation B.7 that this is a general result for any equilibrium and can also be applied to study protein (P) + ligand (L) to protein-ligand (PL) complex equilibria. In this case one finds that,

$$d\ln K = -\left(F_{P,PL} - F_{P,P} - F_{P,L}\right)d\beta + \sum_i (N_{PLi} - N_{Pi} - N_{Li})(d\beta\mu_i - \bar{h}_i d\beta) \tag{B.70}$$

where $K = \rho_{PL}/\rho_P\rho_L$, and we have used the fact that $d\mu_{PL} = d\mu_P + d\mu_L$ to eliminate $d\mu_{PL}$. The summations in the ε_P term and above expression only involve thermodynamically independent species (solvent, protein, ligand, etc). Simulated enthalpy changes are difficult to determine and generally require multiple simulations of the equilibrium constant at different temperatures. The ability to calculate enthalpy changes for these processes from a single computer simulation is particularly attractive.

The expressions provided in Equations B.41 and B.42 also have significant consequences. First, the enthalpy change can be decomposed into a series of terms related to the average intra and intermolecular energies and solvent distribution around each solute form. Further decomposition into van der Waals and electrostatic contributions is also possible without additional approximation. Second, the decomposition is exact and different from typical *ad hoc* approaches.[12,13] Third, one can rewrite Equation B.42 to give,

$$F_{P,D}^{\infty} = < E_{DD} > + < E_{D1} > + < N_1 >_D \left(\frac{< E_{11} >_D}{< N_1 >_D} - H_1^o\right) - < N_1 >_o (E_1^o - H_1^o) \tag{B.71}$$

which indicates that the true measure of the local solvent contribution involves both the number of solvent molecules in the local region, and how their average energy differs from the molar enthalpy of the pure solvent. The neglect of the H_1^o contribution in the calculation of simulated

enthalpy changes would lead to a significant error, even when the difference in the number of solvent molecules between both forms is small (typically 2-5 for most proteins),[17] as the value of H_1^o for common water models is large, -45.6 kJ/mol for SPC/E water,[60] for example. Hence, one cannot just determine the change in the solvent-solvent energy when attempting to determine simulated enthalpy changes. Furthermore, it is unclear to what degree implicit solvent or coarse grained models include or approximate terms involving the number of solvent molecules.

Many of the expressions provided here involve differences between extensive quantities that are then intensive in nature. A prime example is Equation 71 as used in Equation B.41. The last two terms are both extensive (dependent on the reference volume), but their difference is independent of this volume. Furthermore, manipulation of the terms in Equation B.71 has to be performed specifically recognizing that these are grand canonical ensemble averages. For instance, as the reference volume increases one might be tempted to write $\langle E_{11}\rangle_D/\langle N_1\rangle_D$ as E_1^o, implying that the last two terms cancel for large volumes. This is incorrect and it is the change in the last two terms with increasing volume that will cancel leaving the same constant value independent of the reference volume (as long as it is large enough to include all the perturbing effects of the solute). In addition, when decomposing the terms found in Equation 44 for example, one can only isolate the intensive term $< \delta\varepsilon_P\delta\varepsilon_P >_D - < \delta\varepsilon_P\delta\varepsilon_P >_o$, and not the extensive term $< \delta\varepsilon_P\delta\varepsilon_P >_D$ itself, even though the latter (pure solvent) term cancels in Equation B.44.

The energy and enthalpy terms used in the definition of the various ε's involve the total internal energy. The kinetic energy contributions can be removed or ignored in some cases. For example, the kinetic energy contributions to the expression for α_P in Equation 8, and the expressions provided in Equations B.30, B.37, B.41, B.42, B.43, and B.50 all cancel and therefore one could replace E with just the potential energy and ignore the kinetic energy contribution to the partial molar enthalpies. However, this is not the case for the C_P expression provided in Equation 8, or the expressions provided in Equations B.36, B.44, and B.55, where the ideal terms do not cancel, although the contribution (1/2 R per classical degree of freedom) is often small compared to the heat capacity change associated with the equilibrium itself, or will cancel in the case of protein denaturation (n = 1).

When studying protein denaturation it is clear that one is dealing with a transition where the protein changes from a set of relatively few native like conformations to a (potentially) very

large number of denatured or unfolded conformations. However, this does not significantly affect the results presented here. If we consider a collection of denatured forms it can be shown that,

$$d\ln\rho_D = \rho_D^{-1}\sum_i \rho_{D,i} d\ln\rho_{D,i} = \sum_i x_{D,i} d\ln\rho_{D,i} \tag{B.72}$$

where $x_{D,i}$ is the number fraction of denatured form *i*. Therefore, the F's and N_{ij}'s used in Equations 26 and 41 simply become averages over the individual denatured forms weighted by their fractional populations, or a simple time average in a simulation.

In the previous sections we attempted to outline how one could obtain specific contributions to, or correlations with, each form present in the equilibrium. This required additional experimental information. The experimental data can come in a variety of forms and therefore in our previous discussion we just provided expressions for changes in h_2^* with temperature – without invoking a specific process. One process relevant to the present discussion is the process of solvation. In this case the enthalpy change (Δh_2^*) corresponds to the process of transferring a molecule of *i* from a fixed position in an ideal gas phase to a fixed position in the solution at the T, P, and composition of interest. In principle, this information should be amenable to experiment. However, this might not always be true in practice. For instance, the (gas to solution) solvation enthalpy may be available for small volatile solutes, but not for proteins. A more practical application would involve the study of changes in enthalpy between pure water and mixed solvent systems, where the change in enthalpy would then involve the enthalpies of transfer between solvent systems. Derivatives of the protein pseudo enthalpy could be replaced by experimentally available protein heat capacities noting that $c_{p,2}^* = \bar{c}_{p,2} - (\partial RT^2\alpha_P/\partial T)_{P,\{N\}} - 3R/2$, as indicated by Equation 10.

In an effort to establish the relative importance of each term in Equations B.26, B.27, B.41 and B.44 one can examine existing data concerning protein denaturation. Experimental data for the pressure and thermal denaturation of Ribonuclease A at 295 L and pH 2 in D_2O is available.[61,62] The observed difference in volume of -21 cm^3/mol obtained from pressure denaturation is small, especially compared to the native state volume of 9500 cm^3/mol, and corresponds to slightly more than one water molecule (V_1^o = 18 cm^3/mol). The second derivative of the equilibrium constant with respect to pressure is determined to be 6.1 x 10^{-6} bar^{-2}. Using the above data one finds the κ_T^o term in Equation B.27 to contribute a negligible -2 x 10^{-9} bar^{-2},

while the final term contributes 2 x 10^{-6} bar^{-2}. Therefore, the final term and the $< \delta N_1 \delta N_1 >$ term are similar in magnitude. Thermal denaturation data provide a standard enthalpy change and heat capacity of 200 kJ/mol and 4800 J/mol/K, respectively. Hence, the $C_{P,m}^{o}$ term in Equation B.44 is negligible, while the other two terms would appear to be significant.

It is tempting to simplify some of the expressions provided previously. However, this should be performed with care. For instance, the pressure effect can be written by reference to Equation B.10 as,

$$\left(\frac{\partial \ln K}{\partial P}\right)_{\beta,\{N\}} = \beta \sum_i (N_{Ai} - nN_{Mi})\overline{V}_i = -\beta(V_A^* - nV_M^*) \tag{B.73}$$

which seems logical. It might then be tempting to write,

$$\left(\frac{\partial^2 \ln K}{\partial P^2}\right)_{\beta,\{N\}} = -\beta\left[\left(\frac{\partial V_A^*}{\partial P}\right)_{\beta,\{N\}} - n\left(\frac{\partial V_M^*}{\partial P}\right)_{\beta,\{N\}}\right] \tag{B.74}$$

and to use the expression provided by Equation 24 for the two derivatives. This is incorrect and differs from the correct result provided in Equation 20. The reason is that Equation B.24 was developed for a system of thermodynamically independent composition variables and therefore requires that all N (including N_A and N_M) are held constant. This is clearly not the case according to Equation B.16. In contrast, the development of Equation B.20 correctly captures the inherent dependence of N_A (and N_M) on pressure and only assumes that their sum (N_2) is constant. Similar arguments also explain why one cannot simply replace the F_P's in Equation B.30, for instance, with their values suggested by Equation B.8. Furthermore, computer simulations or experiments which determine how the volume of a single protein form varies with pressure will provide information concerning the compressibility of that form. However, the difference in compressibility between these independent forms is not simply the compressibility associated with the chemical equilibrium itself.

The fluctuation based results for the effect of pressure on protein denaturation involve differences between the water distributions surrounding each protein form. This is clearly related to the protein volume. For an infinitely dilute solute (2) in a single solvent (1) the pseudo volume is given by $-N_{21}^{\infty}V_1^{o}$ and can be considered as the ensemble average of a series of protein volumes given by $V - N_1 V_1^{o}$, where N_1 is the number of solvent molecules surrounding a single

protein in the same fixed volume V for each member of the ensemble. In this case the fluctuations in the protein volume are then given by $< \delta N_1 \delta N_1 >_2 (V_1^o)^2$, which is one of the terms that appears in Equations B.27 and B.29. Therefore, fluctuations in the water distribution and fluctuations in the protein volume are intimately related. However, this relationship is only exact when one measures the protein volume using the local number of water molecules in the solution and, unfortunately, this is not the only term that appears in Equations B.27 and B.29.

Experimental information concerning the compressibility of proteins can also be obtained from sound velocity studies.[63-65] This approach has the advantage of probing the compressibility of proteins under normal pressures and temperatures. The resulting isentropic protein compressibilities (κ_S) are, however, much more difficult to interpret both experimentally and theoretically. For instance, the isentropic compressibility of a solution mixture is given by the thermodynamic relationship,

$$RT\kappa_S = RT\kappa_T - RT\alpha_P \left(\frac{\partial T}{\partial P}\right)_{S,\{N\}} = RT\kappa_T - \frac{(RT^2\alpha_P)^2 V_m}{RT^2 C_{P,m}} \tag{B.75}$$

which can be expressed in terms of fluctuating quantities using Equation B.8. Experiments provide partial molar or apparent molar isentropic protein compressibilities. These involve derivatives of the above expression and thereby contain a variety of fluctuating quantities. The analysis is greatly simplified by transforming to partial molar isothermal compressibilities,[65] which can then be analyzed in a more traditional manner (see Appendix 2). Hence, we have not provided the expressions for partial molar isentropic compressibilities here.

Finally, it is important to recognize that changes to the equilibrium constant also involve contributions from the internal partition function, specifically changes to the vibrational modes, especially when the temperature is varied. To illustrate this one can write (from Equation B.9 and B.14) for protein denaturation,

$$\left(\frac{\partial \ln K}{\partial \beta}\right)_{P,\{N\}}^{\infty} = -\left(\frac{\partial \beta(\mu_D^{*,\infty} - \mu_N^{*,\infty})}{\partial \beta}\right)_{P,\{N\}} = -(h_D^{*,\infty} - h_N^{*,\infty}) = -\left(F_{P,D}^{\infty} - F_{P,N}^{\infty}\right) \tag{B.76}$$

The pseudo chemical potential terms contain the internal partition function whose dependence on temperature can be significant. Hence, analysis of experimental data on protein denaturation will explicitly include the vibration contributions in the extracted FP's through the

energy terms. However, simulations of classical systems will not include such contributions, or they will only be included to the degree that the force field has implicitly accounted for such affects during the parameter development.

B.6 Conclusions

Expressions describing how a chemical equilibrium responds to changes in pressure, temperature and composition have been provided in terms of local fluctuations around the relevant chemical forms in solution. The expressions can be used to analyze experimental data regarding any chemical equilibrium which follows Equation B.11 in any multicomponent mixture at any composition, or they can be used to analyze or predict such effects using computer simulation. In particular, we provide exact expressions for determining enthalpy, heat capacity, and compressibility changes associated with a chemical equilibrium from a single simulation. The resulting expressions contain terms which involve particle-energy, energy-energy, and particle-particle correlations for the enthalpy, heat capacity and compressibility, respectively, for processes at infinite dilution in a single solvent. However, additional terms are also present which render the expressions non trivial and different from more intuitive, but more approximate, approaches.

B.7 Acknowledgements

The Authors would like to thank Samantha Weerasinghe and Elizabeth Ploetz for valuable discussion. The project described was supported by Grant Number R01GM079277 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

B.8 Appendix

Appendix 1 – Derivatives of Fluctuating Quantities in the Grand Canonical Ensemble

Here we outline the general approach for obtaining fluctuating quantities which can be used to help develop derivatives of KBIs with respect to pressure, temperature and composition.

The differential for a general intensive fluctuating quantity in the grand canonical ensemble can be written,

$$d<\delta X\delta Y> = -<\delta X\delta Y\delta E>d\beta + <\delta X\delta Y>\mathrm{dln}V + \sum_k <\delta X\delta Y\delta N_k>d\beta\mu_k$$

(B.77)

where we have used the fact that d< δX δY > = < δX δY >/V dV when < δX δY >/V is intensive and therefore independent of V. Taking derivatives of this equation with the appropriate values of X and Y provides the expressions required for Equations B.23, B.38 and B.65. This approach can also be used for other partial molar quantities as illustrated below (Appendix 2). When X and Y are both particle numbers this provides a route to derivatives of the KBIs or other fluctuating quantities (thermal expansion, compressibility, heat capacity) with respect to pressure, temperature, and composition, which could be used to further analyze the properties of solution mixtures. The simplest results are those provided for the B_{ij}'s,

$$\left(\frac{\partial B_{ij}}{\partial P}\right)_{T,\{N\}} = \beta V^{-1}\sum_k < \delta N_i\delta N_j\delta N_k > \bar{V}_k$$

$$\left(\frac{\partial B_{ij}}{\partial \beta}\right)_{P,\{N\}} = -V^{-1} < \delta N_i\delta N_j\delta\varepsilon_P >$$

$$\left(\frac{\partial B_{ij}}{\partial < N_l >}\right)_{P,T,\{N\}'} = V^{-1}\sum_k < \delta N_i\delta N_j\delta N_k > \mu_{kl}$$

(B.78)

where $\mu'_{kl} = \beta(\partial\mu_k/\partial < N_l >)_{P,T,\{N\}'}$.

Appendix 2 – Partial Molar Heat Capacities, Thermal Expansions, and Compressibilities

In this section we derive expressions for several partial molar quantities used in the previous discussion but not available in the literature. The partial molar constant pressure heat capacity can be obtained starting from the definition,

$$RT^2\bar{c}_{p,i} = RT^2\left(\frac{\partial\bar{h}_i}{\partial T}\right)_{P,\{N\}} = \left(\frac{\partial RT^2C_P}{\partial < N_i >}\right)_{T,P,\{N\}'}$$

(B.79)

Therefore, using the expression in Equation 8 for C_P and then Equation B.77 with X = Y = ε_P one finds,

$$RT^2\bar{c}_{P,i} = \rho\bar{V}_i RT^2 C_{P.m} + \sum_j <\delta\varepsilon_P \delta\varepsilon_P \delta N_j > \mu'_{ji}$$

(B.80)

where ρ is the total number density. As the chemical potential derivatives and partial molar volume can also be expressed in terms of fluctuations (KBIs) this is the desired result, although including these additional fluctuations here is not particularly informative. We note that $C_{P,m} = \sum_i x_i \bar{c}_{P,i}$ and hence the double summation over the last term in Equation B.78 must be zero. The same approach can be used for the corresponding constant volume quantities. First, we note that the quantities used in Equation B.54 are given by,

$$RT^2\bar{c}_{V,i} \equiv RT^2\left(\frac{\partial\bar{e}_i}{\partial T}\right)_{V,\{N\}} = \left(\frac{\partial RT^2 C_V}{\partial < N_i >}\right)_{T,P,\{N\}'}$$

(B.81)

in terms of the partial molar energies ($\bar{e}_i$). Using the expression for C_v provided in Equation B.51 leads to,

$$RT^2\bar{c}_{V,i} = \rho\bar{V}_i RT^2 C_{V.m} + \sum_j <\delta\varepsilon_V \delta\varepsilon_V \delta N_j > \mu'_{ji}$$

(B.82)

as a final result.

Expressions for the partial molar thermal expansions are slightly more complicated. First, we note that from our definition one has,

$$RT^2\bar{\alpha}_{p,i} \equiv RT^2\left(\frac{\partial \ln\bar{V}_i}{\partial T}\right)_{P,\{N\}} = RT^2\alpha_P + \frac{V}{\bar{V}_i}\left(\frac{\partial RT^2\alpha_P}{\partial < N_i >}\right)_{T,P,\{N\}'}$$

(B.83)

From Equation B.8 the general expression for the thermal expansion can be written $RT^2\alpha_P = -V^{-1}\sum_j < \delta N_j \delta\varepsilon_P > \bar{V}_j$ and therefore,

$$\left(\frac{\partial RT^2\alpha_P}{\partial < N_i >}\right)_{T,P,\{N\}'} = -\frac{\bar{V}_i}{V}RT^2\alpha_P - \frac{1}{V}\sum_j\left(\frac{\partial(< \delta N_j \delta\varepsilon_P > \bar{V}_j)}{\partial < N_i >}\right)_{T,P,\{N\}'}$$

(B.84)

which, after using Equation B.77 with X = $N_j\bar{V}_j$ and Y = εP, provides the final result,

$$RT^2\bar{\alpha}_{p,i} = RT^2\alpha_P - \frac{1}{\bar{V}_i}\sum_{j,k}\bar{V}_j<\delta N_j\delta N_k\delta\varepsilon_p>\mu'_{ki} \tag{B.85}$$

We note that $\alpha_P = \sum_i \phi_i\bar{\alpha}_{P,i}$ and hence the triple summation over the last term in Equation B.83 must be zero.

Finally, the partial molar isothermal compressibilities can be obtained from our initial definition,

$$RT\bar{\kappa}_{T,i} \equiv RT\left(\frac{\partial \ln\bar{V}_i}{\partial P}\right)_{T,\{N\}} = RT\kappa_T + \frac{V}{\bar{V}_i}\left(\frac{\partial RT\kappa_T}{\partial < N_i >}\right)_{T,P,\{N\}'} \tag{B.86}$$

From Equation B.8 the general expression for the compressibility can be written $RT\kappa_T = V^{-1}\sum_{j,k} < \delta N_j\delta N_k > \bar{V}_j\bar{V}_k$ and therefore,

$$\left(\frac{\partial RT\kappa_T}{\partial < N_i >}\right)_{T,P,\{N\}'} = -\frac{\bar{V}_i}{V}RT\kappa_T + \frac{1}{V}\sum_{j,k}\left(\frac{\partial(< \delta N_j\delta N_k > \bar{V}_j\bar{V}_k)}{\partial < N_i >}\right)_{T,P,\{N\}'} \tag{B.87}$$

which, after using Equation B.77 with X = $N_j\bar{V}_j$ and Y = $N_k\bar{V}_k$, provides the final expression,

$$RT\bar{\kappa}_{T,i} = RT\kappa_T + \frac{1}{\bar{V}_i}\sum_{j,k,l}\bar{V}_j\bar{V}_k<\delta N_j\delta N_k\delta N_l>\mu'_{li} \tag{B.88}$$

We note that $\kappa_T = \sum_i \phi_i\bar{\kappa}_{T,i}$ and hence the quadruple summation over the last term in Equation B.86 must be zero.

Reference List

(1) K. A. Dill, *Biochemistry* **1990**, 29, 7133.

(2) J. F. Brandts, *Journal of American Chemical Society* **1964**, 86, 4291.

(3) A. Zipp and W. Kauzmann, *Biochemistry* **1973**, 12, 4217.

(4) C. Tanford, *Journal of American Chemical Society* **1964**, 86, 2050.

(5) V. A. Parsegian, R. P. Rand, N. L. Fuller, and D. C. Rau, *Methods in Enzymology* **1986**, 127, 400.

(6) F. Meersman, C. M. Dobson, and K. Heremans, *Chemical Society Review* **2006**, 35, 908.

(7) R. Liu, H. Barkhordarian, S. Emadi, C. B. Park, and M. R. Sierks, *Neurobiology of Disease* **2005**, 20, 74.

(8) S. Narayanan and B. Reif, *Biochemistry* **2005**, 44, 1444.

(9) D. K. Klimov, J. E. Straub, and D. Thirumalai, *Proceeding of National Academy of Sciences* U.S.A. **2004**, 101, 14760.

(10) D. R. Canchi, D. Paschek, and A. E. Garcia, *Journal of American Chemical Society* **2010**, 132, 2338.

(11) D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan, and W. Wriggers, *Science* **2010**, 330, 341.

(12) P. Setny, R. Baron, and J. A. McCammon, *Journal of Chemical Theory Computation* **2010**, 6, 2866.

(13) L. V. Schafer, D. H. de Jong, A. Holt, A. J. Rzepiela, A. H. de Vries, B. Poolman, J. A. Killian, and S. J. Marrink, *Proceeding of National Academy of Sciences* **2011**, 108, 1343.

(14) T. Lazaridis and M. Karplus, *Biophysical Chemistry* **1999**, 78, 207.

(15) N. V. Prabhu and K. A. Sharp, *Annual Review of Physical Chemistry* **2005**, 56, 521.

(16) S. W. Rick, *Journal of Physical Chemistry B* **2000**, 104, 6884.

(17) C. A. Royer, *Biochim. Biophys. Acta-Protein Structure and Molecular Enzymology* **2002**, 1595, 201.

(18) V. M. Dadarlat and C. B. Post, *Journal of Physical Chemistry B* **2001**, 105, 715.

(19) E. Paci, Biochim. *Biophys. Acta-Protein Structure and Molecular Enzymology* **2002**, 1595, 185.

(20) V. M. Dadarlat and C. B. Post, *Biophysics Journal* **2006**, 91, 4544.

(21) A. Cooper, *Proceeding of National Academy of Sciences* U.S.A. **1976**, 73, 2740.

(22) C. Scharnagl, M. Reif, and J. Friedrich, *Biochim. Biophys. Acta-Proteins Proteomics* **2005**, 1749, 187.

(23) A.Wallqvist, D. G. Covell, and D. Thirumalai, *Journal of American Chemical Society* **1998**, 120, 427.

(24) B. J. Bennion and V. Daggett, *Proceeding of National Academy of Sciences* U.S.A. **2003**, 100, 5142.

(25) P. J. Rossky, *Proceeding of National Academy of Sciences* **2008**, 105, 16825.

(26) L. Hua, R. H. Zhou, D. Thirumalai, and B. J. Berne, *Proceeding of National Academy of Sciences* **2008**, 105, 16928.

(27) V. Pierce, M. Kang, M. Aburi, S. Weerasinghe, and P. E. Smith, *Cell Biochem. Biophys*. **2008**, 50, 1.

(28) S. Shimizu, *Proceeding of National Academy of Sciences* **2004**, 101, 1195.

(29) P. E. Smith, *Journal of Physical Chemistry B* **2004**, 108, 18716.

(30) I. L. Shulgin and E. Ruckenstein, *Journal of Chemical Physics* **2005**, 123, 054909.

(31) E. Ruckenstein and I. L. Shulgin, *Adv. Colloid Interface Sci.* **2006**, 123, 97.

(32) I. L. Shulgin and E. Ruckenstein, *Biophys. J.* **2006**, 90, 704.

(33) J. Rosgen, B. M. Pettitt, and D. W. Bolen, *Protein Sci.* **2007**, 16, 733.

(34) J. G. Kirkwood and F. P. Buff, *Journal of Chemical Physics* **1951**, 19, 774.

(35) J. P. O'Connell, *Mol. Phys.* **1971**, 20, 27.

(36) A. Ben-Naim, *Molecular Theory of Solutions.* Oxford University Press, New York, **2006**.

(37) E. Matteoli and G. A. Mansoori, *Fluctuation Theory of Mixtures.* Taylor & Francis, New York, **1990**.

(38) F. P. Buff and R. Brout, *Journal of Chemical Physics* **1955**, 23, 458.

(39) P. G. Debenedetti, *Journal of Chemical Physics* **1987**, 87, 1256.

(40) P. G. Debenedetti, *Journal of Chemical Physics* **1987**, 86, 7126.

(41) P. G. Debenedetti, *Journal of Chemical Physics* **1988**, 88, 2681.

(42) P. G. Debenedetti, *Molecular Simu*lation **1989**, 2, 33.

(43) E. A. Ploetz and P. E. Smith, "Local fluctuations in solution mixtures," *Journal of Chemical Physics* (submitted).

(44) S. Shimizu and C. L. Boon, *Journal of Chemical Physics* **2004**, 121, 9147.

(45) M. Aburi and P. E. Smith, *Journal of Chemical Physics B* **2004**, 108, 7382.

(46) R. L. Perry, J. C. Telotte, and J. P. O'Connell, *Fluid Phase Equilibria* **1981**, 5, 245.

(47) R. L. Perry and J. P. O'Connell, *Molecular Physics* **1984**, 52, 137.

(48) A. Ben-Naim, *Journal of Chemical Physics* **1975**, 63, 2064.

(49) M. B. Gee and P. E. Smith, *Journal of Chemical Physics* **2009**, 131, 165101.

(50) K. E. Newman, *Chemical Society Review* **1994**, 23, 31.

(51) M. Kang and P. E. Smith, *Journal of Chemical Physics* **2008**, 128, 244511.

(52) P. E. Smith, *Journal of Chemical Physics* **2008**, 129, 124509.

(53) D. J. Adams, *Molecular Phy*sics **1975**, 29, 307.

(54) T. Imai and Y. Sugita, *Journal of Physical Chemistry B* **2010**, 114, 2281.

(55) P. E. Smith, *Journal Physical Chemical B* **2006**, 110, 2862.

(56) R. F. Greene Jr and C. N. Pace, *Journal of Biological Chemistry* **1974**, 249, 5388.

(57) G. I. Makhatadze, *Journal of Physical Chemistry B* **1999**, 103, 4781.

(58) P. E. Pjura, M. E. Paulaitis, and A. M. Lenhoff, *AIChE Journal* **1995**, 41, 1005.

(59) S. Weerasinghe and P. E. Smith (unpublished).

(60) H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *Journal of Physical Chemistry* **1987**, 91, 6269.

(61) G. I. Makhatadze, G. M. Clore, and A. M. Gronenborn, *National Structure Biology* **1995**, 2, 852.

(62) K. E. Prehoda, E. S. Mooberry, and J. L. Markley, *Biochemistry* **1998**, 37, 5785.

(63) H. Shiio, T. Ogawa, and H. Yoshihashi, *Journal of American Chemical Society* **1955**, 77, 4980.

(64) K. Gekko and H. Noguchi, *Journal of Physical Chemistry* **1979**, 83, 2706.

(65) K. Gekko and Y. Hasegawa, *Biochemistry* **1986**, 25, 6563.

Appendix C - Copy of the Permission Letter from the Publisher

AMERICAN INSTITUTE OF PHYSICS LICENSE TERMS AND CONDITIONS

Mar 06, 2012

License Number	2863210860542
Order Date	Mar 06, 2012
Publisher	American Institute of Physics
Publication	Journal of Chemical Physics
Article Title	Fluctuation theory of molecular association and conformational equilibria
Author	Yuanfang Jiao, Paul E. Smith
Online Publication Date	Jul 5, 2011
Volume number	135
Issue number	1
Type of Use	Thesis/Dissertation
Requestor type	Author (original article)
Format	Print and electronic
Portion	Excerpt (> 800 words)
Will you be translating?	No
Title of your thesis / dissertation	THE DEVELOPMENT OF ACCURATE FORCE FIELDS FOR PROTEIN SIMULATION
Expected completion date	Jun 2012
Estimated size (number of pages)	120
Total	0.00 USD

Terms and Conditions

American Institute of Physics -- Terms and Conditions: Permissions Uses

American Institute of Physics ("AIP") hereby grants to you the non-exclusive right and

State of New York. Both parties hereby submit to the jurisdiction of the courts of New York County for purposes of resolving any disputes that may arise hereunder.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500733848.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.